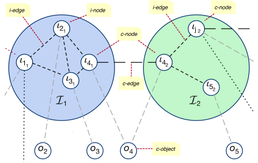
Paolo Lo Giudice

was born in Reggio Calabria, Italy, in 1991.
He received the MSc degree in ICT Engineering from the University Mediterranea of Reggio Calabria in October 2016.
His research interests include Social Network Analysis, Social Internetworking, Source and Data Integration, Internet of Things, Innovation Management, Knowledge Extraction and Representation, Biomedical applications and Data Lake.
He is also co-authored of some pubblications in peer-reviewed national and international journals and several conference contributions.

DOTTORATO di RICERCA in INGEGNERIA dell'INFORMAZIONE
SCUOLA di DOTTORATO
Università degli Studi *Mediterranea* di Reggio Calabria



In the big data era, the number, the volume and the variety of available data sources are dramatically increasing. As a consequence, one of the main open issues to address in computer science research consists of uniformly extracting knowledge and facing decision problems in heterogeneous application contexts. However, as generally happens, a solved problem becomes an opportunity. In fact, if we were able to define a model suitable to uniformly represent and handle highly heterogeneous data formats, we could use it to manage data coming from several research contexts. In other words, an approach designed to solve an open problem in one context can be easily transposed to address other open issues in other contexts. This thesis aims at providing a contribution in this setting. Indeed, it proposes a social network-based approach to uniformly extract knowledge and support decision making concerning disparate research contexts. In particular we will focus on four contexts: Biomedical Engineering, Data Lakes, Internet of Things and Innovation Management

COLLANA DELLA SCUOLA DI DOTTORATO DELL'UNIVERSITA' DEGLI STUDI *MEDITERRANEA* DI REGGIO CALABRIA

Paolo LO GIUDICE

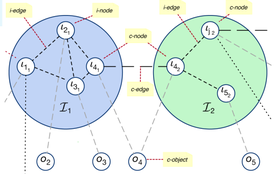
A network-based approach to uniformly extract knowledge and support decision making in heterogeneous application context

ISBN: 978-88-99352-40-0

Progetto grafico
Centro Stampa d'Ateneo

Paolo LO GIUDICE

A network-based approach to uniformly extract knowledge and support decision making in heterogeneous application context



Supervisor: Prof. Domenico URSINO
Coordinator: Prof. Tommaso ISERNIA
S.S.D. ING-INF/05
XXXII Ciclo

Collana
Quaderni del Dottorato di Ricerca in Ingegneria dell'Informazione
Quaderno n°45

DOCTORAL SCHOOL
UNIVERSITA' *MEDITERRANEA* DI REGGIO CALABRIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE, DELLE INFRASTRUTTURE E DELL'ENERGIA SOSTENIBILE (DIIES)

PHD IN
INFORMATION ENGINEERING

S.S.D. ING-INF/05
XXXII CICLO

A network-based approach to uniformly extract knowledge and support decision making in heterogeneous application contexts

CANDIDATE
Paolo LO GIUDICE

ADVISOR
Prof. Domenico URSINO

COORDINATOR
Prof. Tommaso ISERNIA

REGGIO CALABRIA, January 2020

Finito di stampare nel mese di **Gennaio 2020**

Edizione CSdA Centro Stampa d'Ateneo

Quaderno N. 45
Collana *Quaderni del Dottorato di Ricerca in Ingegneria dell'Informazione*
Curatore *Prof. Tommaso Isernia*

ISBN 978-88-99352-40-0

Università degli Studi *Mediterranea* di Reggio Calabria
Salita Melissari, Feo di Vito, Reggio Calabria

PAOLO LO GIUDICE

A network-based approach to uniformly extract knowledge and support decision making in heterogeneous application contexts

The Teaching Staff of the PhD course in
INFORMATION ENGINEERING
consists of:

Tommaso ISERNIA (coordinator)
Giovanni ANGIULLI
Pier Luigi ANTONUCCI
Giuseppe ARANITI
Antoine BERTHET
Francesco BUCCAFURRI
Claudia CAMPOLO
Rosario CARBONE
Riccardo CAROTENUTO
Luigi CELONA
Salvatore COCO
Mariantonia COTRONEI
Lorenzo CROCCO
Dominique DALLET
Claudio DE CAPUA
Francesco DELLA CORTE
Giuliana FAGGIO
Pasquale FILIANOTI
Patrizia FRONTERA
Sofia GIUFFRE'
Voicu GROZA
Antonio IERA
Gianluca LAX
Aime' LAY EKUAKILLE
Giacomo MESSINA
Antonella MOLINARO
Andrea MORABITO
Rosario MORELLO
Fortunato PEZZIMENTI
Sandro RAO
Ivo RENDINA
Domenico ROSACI
Giuseppe RUGGERI
Mariateresa RUSSO
Valerio SCORDAMAGLIA
Domenico URSINO

Significa qualcosa che appartiene al passato,
ad un futuro immaginato che però non c'è

Contents

Part III Internet of Things

List of Figures

List of Tables

1

Introduction

1.1 Motivations

In the big data era, the number, the volume and the variety of available data sources are dramatically increasing. As a consequence, one of the main open issues to address in computer science research consists of uniformly extracting knowledge and facing (very complex) decision problems in heterogeneous application contexts. However, as generally happens, a solved problem becomes an opportunity. In fact, if we were able to define a model suitable to uniformly represent and handle highly heterogeneous data formats, we could use it to manage data coming from several research contexts. In other words, an approach designed to solve an open problem in one context can be easily transposed to address other open issues in other contexts. This thesis aims at providing a contribution in this setting. Indeed, it proposes a social network-based approach to uniformly extract knowledge and support decision making concerning disparate research contexts. In particular we will focus on four contexts, namely: Biomedical Engineering (BE - specifically electroencephalogram tracks to investigate neurological disorders), Data Lakes (DL), Internet of Things (IoT) and Innovation Management (IM - specifically patent data to investigate innovation trends).

The attempt to uniformly handle data sources characterized by heterogeneous formats for extracting knowledge and supporting decision making has been performed in the past, when most of available data were structured or semi-structured [288, 60, 62, 123, 348, 351]. However, with the advent of the big data phenomenon, most of available data (i.e., about 80%) are unstructured [110]. This is rapidly changing the coordinates of several research fields. So, the need of new models and approaches to handle data with disparate formats is compulsory. As for this exigency, it was shown that network-based models and approaches have the flexibility and, at the same time, the power of effectively and efficiently handling data represented in heterogeneous formats [76]. For this purpose, the advances in the Operations Research (OR) field,

especially in Graph Optimization (GO), which network-based models and approaches derive from, can successfully support knowledge extraction and decision making.

For instance Social Network Analysis (SNA) has been extensively investigated from some decades and, with the advent of Online Social Networks (OSNs), it has become one of the hot topics in computer science. In this context, several interesting results concerning information diffusion [301], homophily [305], centrality [162], crawling [84], etc., have been already found. Network models have also been successfully adopted to face issues concerning IoTs [49], with particular reference to Wireless Sensor Networks (WSNs) and event and anomaly detection [114, 165]. Most of these studies focus on the analysis of data produced by single devices [365], while few are based on the processing of aggregated data acquired by WSNs [90]. Here, network based models have been mainly applied to WSN design and routing [287, 415, 15, 57, 197, 187]. The usage of these models in Biomedical Engineering has been successfully experimented in the past to handle electroencephalographic (EEG) and electrocardiographic (ECG) data [279]. On the other side, brain diseases have been largely analyzed in Biomedical Engineering. Here, EEG analysis supports the study of problems related to the brain, in a non-invasive and economic fashion. In this context, network based models have been used for the diagnosis of several pathological states in humans [192, 478, 389]. Finally, the same models have been already used to face several problems concerning Innovation Management. Among them, we cite the detection of hub institutions in a country [153].

In this thesis, we will examine the network-based models presented in the past literature to represent structured and semi-structured sources [76]. In particular, we will determine the pros and the cons of each of them. Furthermore, we will investigate the features they need to have for handling unstructured data. Finally, we will define a new model maintaining the pros and avoiding the cons of the previous ones by adding the necessary features to make it capable of handling also unstructured data.

In the same way, we can define a unique network analysis-based approach for extracting knowledge and supporting decision making in disparate contexts. Starting from the past literature, we will define new and more appropriate techniques for extracting knowledge and supporting decision making in several domains. This way of proceeding will return a set of general techniques, well suited for the new model and that, when applied to a certain context, allow us to address issues typical of that context. In other words, we will produce a set of generic and, at the same time, powerful template techniques, which can be specialized in many application fields and can support the resolution of problems typical of each of these fields.

We will apply the new model and the new approach to four contexts of interest, namely: *(i)* Neurological Disorders, *(ii)* Data Lakes, *(iii)* IoT, and *(iv)* Information Management.

In the whole thesis, we will underline the commonalities of the models and approaches described in the four contexts. In particular, we will try to define some best practices and we specify some guidelines for modifying models in order to further empower them for future research efforts.

1.1.1 Neurological Disorders

The first context refers to EEG data. Here, we propose a new network-based approach to help experts to investigate neurological disorders in which the connections among brain areas play a key role. Our approach receives the EEG of a patient and associates a network with it, with nodes that represent electrodes and with edges that denote the disconnection degree of the corresponding brain areas. Starting from this network, we investigate the strength of the connections between brain areas and use this strength to investigate three neurological disorders, namely Creutzfeldt-Jacob Disease (CJD), Alzheimers Disease (AD) and Childhood Absence Epilepsy (CAE).

In recent years, the incidence of neurological disorders is growing also because population is aging in most countries. At the same time, the efforts to design approaches capable of determining the onset of these disorders and of monitoring their course in the corresponding patients are intensifying [138, 207, 463]. Even the tools supporting neurologists in their activities are becoming more complex and sophisticated (think, for instance, of electroencephalograms with 256 electrodes, instead of the classical ones with 19 electrodes). The counterpart of these important advances is the need of handling huge amounts of data that experts have difficulty to analyze manually. In this scenario, automatic tools helping experts to analyze available data are becoming mandatory.

Among the many diagnostic tools available to neurologists, electroencephalogram (hereafter, EEG) is one of the least invasive. For this reason, it is adopted to support the analyses of many neurological disorders. In the literature, many techniques to process EEG data have been proposed, and most of them are based on signal analysis [88, 220, 343, 418, 456, 422].

An EEG can be easily modeled as a network. Indeed, several approaches that use networks to model EEGs and to investigate neurological disorders have been presented [122, 280, 295, 370, 416, 472] in the past. After having modeled an EEG as a network, these approaches generally use basic concepts and metrics of network analysis (e.g., centrality measures, diameter, path length) to help an expert in her diagnosis.

It is well known that, in many neurological investigations, the key role is played by the connections between the brain areas. Network analysis provides some basic parameters to evaluate the connection level of a network. The most known of them are network density and clustering coefficient. However, these two parameters have not been specifically conceived for measuring the connection degree of a network. As a consequence, a challenging issue could be defining a parameter specifically thought for this purpose. Hopefully, this parameter could work better than density and clustering coefficient for evaluating the connection degree of a network. To define it, we observe that cliques play a central role in identifying highly-connected portions of a network. Thus, they could represent the key concept in this task, because the higher the number and the dimension of available cliques in a network and the higher the corresponding connection level.

However, a network associated with an EEG is totally connected, since a voltage difference can be evaluated for each pair of its electrodes. On the other side, voltage difference between two electrodes is an indicator of the strength of the connection between them and, ultimately, between the corresponding brain areas. As a consequence, it is reasonable to use a metric derived from it to weigh the corresponding edges in the network. This metric could represent the distance, or the disconnection level, of the associated brain areas. These edge weights could guide the analyses of the network and, ultimately, of the corresponding patient.

In this scenario, a metric that, starting from the voltage differences, can determine the disconnection level between two nodes is particularly important. We have decided to propose a new approach orthogonal to the metric adopted to weigh network edges.

1.1.2 Data Lakes

The second context will focus on Data Lakes. In particular, we propose a new network-based approach to uniformly manage heterogeneous data lake sources. This approach first models involved sources by means of networks, then it exploits network-based techniques to extract interchema properties and knowledge patterns from them. The extracted knowledge will represent the metadata that are, in turn, the core of a data lake.

Metadata have always played a key role in favoring the cooperation of heterogeneous data sources. This role has become much more crucial with the advent of data lakes, in which case metadata represent the only possibility to guarantee an effective and efficient management of data source interoperability. For this reason, the necessity to define new models and paradigms for metadata representation and management appears crucial in the data lake scenario. We aim at addressing this issue by proposing a new metadata model, well suited for data lakes. Furthermore, to give an idea of

its capabilities, we present an approach that leverages it to "structure" unstructured sources and to extract thematic views from heterogeneous data lake sources.

In the last few years, the "big data phenomenon" is rapidly changing the research and technological "coordinates" of the information system area [93, 451]. For instance, it is well known that data warehouses, generally handling structured and semi-structured data offline, are too complex and rigid to manage the wide amount and variety of rapidly evolving data sources of interest for a given organization, and the usage of more agile and flexible structures appears compulsory [128]. Data lakes are one of the most promising answers to this exigency. Differently from a data warehouse, a data lake uses a flat architecture (so that the insertion and the removal of a source can be easily performed). However, the agile and effective management of data stored therein is guaranteed by the presence of a rich set of extended metadata. These allow a very agile and easily configurable usage of the data stored in the data lake. For instance, if a given application requires the querying of some data sources, one could process available metadata to determine the portion of the involved data lake to examine.

In this scenario, we propose a new metadata model well suited for data lakes. Our model starts from the considerations and the ideas proposed by data lake companies (in particular, it starts from the general metadata classification also used by Zaloni [341]). However, it complements them with new ideas and, in particular, with the power guaranteed by a network-based and semantics-driven representation of metadata. Through this approach, our model can take advantage of all the results already found in network theory and semantic-based approaches. As a result, it can allow a large number of sophisticated tasks that currently adopted metadata models cannot guarantee. For example, it allows the definition of a structure for unstructured data. It also allows the extraction of thematic views from data sources, i.e. the construction of views on one or more topics of interest to the user, obtained by extracting and merging data from different sources.

1.1.3 Internet of Things

The third context regards an IoT scenario. In this case, we will use Social Network Analysis to represent multiple networks of smart objects interconnected to each other through cross objects. Then, we will use this representation to extract knowledge from heterogeneous sensor data streams and to build virtual IoTs in a Multiple IoTs scenario.

The Internet of Things (IoT) is currently considered the new frontier of the Internet, and a lot of research results about this topic can be found in literature. One of the most effective ways to investigate and implement IoT is based on the use of

the social network paradigm: Social Internet of Things (SIoT) is an excellent attempt in this direction. In the last years, social network researchers have introduced new paradigms capable of capturing the growing complexity of this scenario. One of the most known is the Social Internetworking System, which models a scenario comprising several related social networks. We investigate the possibility of applying the ideas underlying Social Internetworking System to IoT, and we propose a new paradigm, called MIoT (Multiple Internets of Things), capable of modelling and handling the increasing complexity of this last context.

MIoT can be seen as an evolution of SIoT (Social Internet of Things). In SIoT, things are empowered with social skills, making them more similar to people [39, 42]. In particular, they can be linked by five kinds of relationship, namely: *(i)* parental object relationship; *(ii)* co-location object relationship; *(iii)* co-work object relationship; *(iv)* ownership object relationship; *(v)* social object relationship. If: *(i)* a node is associated with each thing, *(ii)* an edge is associated with each relationship between things, and, finally, *(iii)* all the nodes and the edges linked by the same relationship are seen as joined together, SIoT can be modeled as a set of five pre-defined networks. Here, some nodes belong to only one network (we call them inner-nodes), whereas other ones belong to more networks (we call them cross-nodes).

The idea underlying SIoT is extremely interesting and, as a matter of fact, has received, and is still receiving, a lot of attention in the literature. However, we think that the number of relationships that might connect things could be much higher than five, and relationships could be much more variegate than the ones currently considered by SIoT. As a consequence, we think that a new paradigm, taking into account this fact, is in order.

We think that the key concepts of SIS can also be applied to things (instead of users) and to relationships between things and we propose the MIoT (Multiple Internets of Things) paradigm. The core of the SIS paradigm is modeling users and their relationships as a unique big network and, at the same time, as a set of related social networks connected to each other thanks to those users joining more than one social network. The MIoT paradigm arises in this scenario. Roughly speaking, a MIoT can be seen as a set of things connected to each other by relationships of any kind and, at the same time, as a set of related IoTs, one for each kind of relationship. Actually, as will be clear in the following, a more precise definition of MIoT would require the introduction of the concept of instance of a thing in an IoT.

1.1.4 Innovation Management

Finally, the fourth context concerns patent data. In this case, we found some inspirations from the approaches that use network-based models to determine institutions

acting as hubs [153]. The ultimate goal is the extraction of knowledge concerning patents, their characteristics and their applicants, as well as information about the influence and the scope of a patent on the other ones.

Patents and collaboration between researchers and, more in general, scientometrics and bibliometrics have been largely investigated in the past. The impressive development of innovations in all the R&D fields and the data available for investigations are growing at a very rapid rate. This has made the adoption of big data centered-techniques compulsory for their analysis. As a matter of fact, the problem of extracting useful knowledge from these data can be seen as a Data Mining problem. In this context, network analysis-based approaches are extremely promising. This is due to the fact that in recent years it has become incredibly important to evaluate the performances of researchers, universities, institutions, etc. Indeed, research collaborations across institutions, firms and countries have been largely investigated in strategy and management literature [409, 308, 79]. Moreover, different studies have been performed to understand whether international flows from developed countries to developing and less-developed ones have some positive effects in these last ones [178]. Furthermore, many studies investigate the impact and the effects of international knowledge flows by focusing on R&D collaborations and inventions and on their impact on innovation [260, 300, 163].

As we pointed out, Social Network Analysis [458, 53, 52, 24, 106, 107, 258, 328] and, more in general, graph theory, have been a prominent family of approaches adopted in the past in this context (see, for istance [276, 36, 46, 72, 446, 359, 10, 13, 277, 237, 103, 71, 11]). Furthermore, it is possible to foresee that they will be even more employed in the future, due to the increasing number of proposals someway involving them.

As it will be clear in the following, our approach presents several features that characterize it with respect to the related ones already proposed in the past. It does not focus on a case study (for instance, on a group of countries). By contrast, it consists of a general methodology for the extraction of several knowledge patterns about innovation geography that can be applied on any country of interest for the user. This is obtained by investigating inventors and not applicants. Furthermore, our approach redefines several metrics, which have been already introduced in SNA or in other research fields, in such a way as to make them suitable to the application context of our interest. It also redefines the concepts of neighborhood, internal neighborhood and external neighborhood of an inventor, which have been previously introduced in totally different research fields. As for this contexts, as we will see in the following, we are able to introduce several new concepts. Finally, our approach defines new metrics about patent and inventor relationships not present in the past; think, for instance,

of the aggregation coefficient and some parameters based on the modified Herfindahl Index for the computation of the heterogeneity of the external collaborations of a country and of the variability of the IPC classes, which the patents of a country refer to.

1.2 Complex Networks as a unifying model for heterogeneous contexts

In this section we provide an overview of a complex network-based model capable of representing disparate scenarios. This model, whose specification will be presented in the next chapters, represents the base for the unifying approach to knowledge extraction that we examine in Section 1.3.

Our complex network-based approach can represent any scenario consisting of several entities (generally of the same type) that interact with each other and that are linked by one or more forms of relationships. Formally speaking, it can be represented as a network:

$$\mathcal{N} = \langle V, E \rangle$$

Here, V is the set of nodes of $\mathcal{N}$. Each node $v_i \in V$ corresponds to an entity, for instancve to an electrode, a metadata label, an object or a patent.

E is the set of the edges of $\mathcal{N}$. Each edge e_{ij} connects the nodes v_i and v_j and can be represented as:

$$e_{ij} = (v_i, v_j, w_{ij})$$

Edges might be weighted. The weight w_{ij} is a measure of the connections strength between v_i and v_j. It is an indicator of the connection/disconnection level of v_i and v_j. Taking into account the peculiarities of the different areas in which the model can operate, we have made our model orthogonal to the different distance measurements that can be used. Indeed, in our experiments, we will employ different types of weight. In some cases, the weight is part of the input (e.g. the PDI in the EEG), while, in other cases, it is computed by pre-processing the input data (think, for instance, of similarity weights in the analysis of similarities between the metadata of different data lakes).

In many of the cases that we have considered, in order to specifically address the analysis of the problems of our interest, we had to build projections of the networks, for instance by removing the edges. This allowed us to make our model more “user-friendly” and “expressive” and, at the same time, more capable of discriminating strong and weak connections between the different network areas.

A network $\mathcal{N}_\pi$, being a projection of a network $\mathcal{N}$, is obtained from this last one by removing the edges with an "excessive" weight and by coloring the others based on their weight. As a matter of fact, if the edges weights represent distance, the edges with an "excessive" weight identify weak connections between the corresponding nodes and can be removed. The remaining edges can be, instead, colored based on their strength. In particular, blue edges denote strong connections, red edges represent intermediate ones and, finally, green edges indicate weak connections. We formalize the network $\mathcal{N}_\pi$ as follows:

$$\mathcal{N}_\pi = \langle V, E_\pi \rangle$$

Here, the nodes of $\mathcal{N}_\pi$ are the same as the ones of $\mathcal{N}$. To define E_π, we consider the distribution of the weights of the edges of $\mathcal{N}$. Specifically, let max_E (resp., min_E) be the maximum (resp., minimum) weight of an edge of E. Starting from max_E and min_E, it is possible to define a parameter $step_E = \frac{max_E - min_E}{10}$, which represents the length of a "step" of the interval between min_E and max_E. We can define $d^k(E)$, $0 \leq k \leq 9$, as the number of the edges of E with weights that belong to the interval between $min_E + k \cdot step_E$ and $min_E + (k+1) \cdot step_E$. All these intervals are closed on the left and open on the right, except for the last one that is closed both on the left and on the right. E_π consists of all the edges of E belonging to $d^k(E)$, where $k \leq th_{max}$.

Now, we can "color" the edges composing E_π. Specifically, $E_\pi = E_\pi^b \cup E_\pi^r \cup E_\pi^g$. Here:

- $E_\pi^b = \left\{ e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{min} \leq k \leq th_{br}} d^k(E) \right\}$;
- $E_\pi^r = \left\{ e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{br} < k \leq th_{rg}} d^k(E) \right\}$;
- $E_\pi^g = \left\{ e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{rg} < k \leq th_{max}} d^k(E) \right\}$.

As will be clear in the following, the projection technique described above, and therefore the corresponding network $\mathcal{N}_\pi$, represent powerful tools at disposal for defining a uniform approach handling knowledge in disparate contexts.

1.3 Social Network Analysis as a unifying approach to knowledge extraction

After having seen that complex networks can represent a unifying model to represent disparate contexts, in this section we aim ad providing a highlight of how, after having modeled contexts as complex or social networks, Social Network Analysis can be exploited as a unifying approach to extract knowledge regarding these contexts. In particular, in the following of this section, in order to prove an overview of this

claim, we will focus on some Social Network Analysis concepts and operators. In the next chapters, we will see that these concepts and operators could be much more numerous.

1.3.1 Clique

In Social Network analysis, one of the most important (and, at the same time, simple and basic) tools for investigating network connection is the concept of clique. We recall that, given a network, a clique of dimension k represents a totally connected subnetwork with k nodes. That is, its induced subgraph is complete. The task of finding whether there is a clique of a given size in a graph (the clique problem) is NP-complete, but, despite this hardness result, many algorithms for finding cliques have been studied.

As for the four contexts examined in this thesis, the concept of clique can be used in different ways. For example, cliques can be adopted as supporting data structures in the process of identifying motifs (i.e., recurring connection patterns within the network), but also to calculate a connection coefficient that, quantitatively, is able to return information about the strength of the network itself.

The most important applications of this concept within our work are the following:

- in neurological disorders, to punctual monitor which areas of the brain are most connected, and, therefore, what parts of the brain continue to operate correctly;
- in innovation management, to identify if there are groups of authors or organizations that often operate together to realize patents.

1.3.2 Centralities

Centrality is one of the most investigated issues in network analysis. It aims at measuring the importance of a node in a network. It allows experts: *(i)* to measure the relevance and the criticity of nodes in their networks; *(ii)* to define forms of distance between network nodes or areas; *(iii)* to measure the cohesion degree of a subnetwork; *(iv)* to identify cohesive subnetworks or network communities.

In the past, several centrality measures have been proposed in the literature [94, 386, 162, 186, 161, 423, 80]. Among them, the most general and best known ones are: *(i) degree centrality*, based on the number of arcs incoming in, or outgoing from, each node; *(ii) closeness centrality*, based on distances between nodes; *(iii) betweenness centrality*, based on the shortest paths connecting pairs of nodes; *(iv) eigenvector centrality*, based on both the number and the centrality of nodes whose outgoing arcs are incident on the nodes of interest.

All these measures, as well as the other ones proposed in the literature, could be adopted in the investigation of the four contexts of interest for this thesis. In particular, they have been adopted as the starting point of several information extraction tasks. Specifically:

- in the IoT analysis, they allowed us to identify cross-nodes within MIoT;
- in the Data Lakes analysis, they allowed us to re-construct a structure for unstructured sources;
- in the innovation management analysis, they allowed us to identify the so-called innovation hubs, i.e., people and/or organization that can favor the development of the whole neighborhood connected with them.

1.3.3 Homophily, Ego Networks and Neighborhoods

Due to the concept of homophily [305] in Social Network Analysis, the behavior of an individual is strictly connected to the one of the individuals most strictly connected to her. Describing and indexing the variation across nodes in the way they are embedded in "local" social structures is the goal of the analysis of ego networks. Ego is an individual focal node. A network has as many egos as it has nodes. Egos can be persons, groups, organizations, or whole societies. A neighborhood, instead, is the collection of an ego and all the nodes to whom it has a connection at some path length. In Social Network Analysis, the "neighborhood" involved in ego networks is almost always one-step; that is, it includes only the ego and the actors that are directly adjacent to her. The neighborhood also includes all of the ties among all of the actors to whom ego has a direct connection. The boundaries of ego networks are defined in terms of neighborhoods.

Homophily, ego networks and neighbors have been extensively exploited throughout this thesis and allowed us to extract knowledge in all the four contexts of our interest. In particular:

- in the IoT analysis, they allowed us to identify communities within object networks;
- in the Data Lakes analysis, they represent the starting point of the Knowledge Pattern extraction process;
- in the innovation management analysis, they allowed us to determine hubs and, more in general, the influence of a patent, an authors or an organization on the connected ones;
- in neurological disorders, to investigate the connection level of a portion of brain.

1.4 A sketch of possible applications

In this section, we provide a sketch of possibile applications of our approach. The details about this issue can be found in the next chapters of this thesis.

1.4.1 Neurological Disorders

The EEGs to perform our investigation were provided by different Italian centers (i.e., University "Magna Graecia" of Catanzaro, Neurologic Institute "Carlo Besta" of Milano, Istituto Bonino-Pulejo and Neurologic Institute of the University of Catania). They regard a group of patients with neurological disorders (in particular, patients suffering from Creutzfeldt-Jacob Disease - CJD -, Mild Cognitive Impairment - MCI -, Alzheimer's Disease - AD -, Childhood Absence Epilepsy - CAE) examined in the last 15 years in these centers. The EEGs were recorded through scalp electrodes placed according to the international 10-20 system. The specific montage was: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, G2, Cz and Pz, where G2 (located between electrodes Fz and Cz) was used as reference. Each EEG was recorded in the morning in a comfortable, eye closed, resting state. The mean recording length is 20 min. The EEG was high-pass filtered at 0.5 Hz, low-pass filtered at 70 Hz. Furthermore a 50 Hz notch filter was also applied. The EEG traces were then downsampled to 256 Hz. The artifactual epochs were visually detected and marked by the EEG experts and later excluded from the analysis. Cleaned data was stored in a MongoDB database [6]. To give an idea of it, we report some of its features: *(i)* size = 357.8 MB; *(ii)* number of collections = 10 ; *(iii)* number of objects = 20; *(iv)* number of indexes = 10; *(v)* index size = 160 KB.

Clearly, we investigated the three neurological disorders separately. Some important results we have found are the following:

- We have defined an approach to identify and characterize Periodic Sharp Wave Complexes in EEGs. These are important indicators of the presence of CJD in a patient.
- We have defined an approach able to discriminate patients who convert from MCI to AD and also to predict, at least partially, the occurrence of an epileptic seizure.
- We have defined an approach able to predict, at least partially, the occurrence of an epileptic seizure.

1.4.2 Data Lakes

To perform our experiments about data lakes, we constructed a set DS of data sources consisting of 2 structured sources, 4 semi-structured ones (2 of which were XML

sources and 2 were JSON ones), and 4 unstructured ones (2 of which were books and 2 were videos). All these sources stored data about environment and pollution. To describe unstructured sources, we considered a list of keywords for each of them. These keywords were derived from Google Books, for books, and from YouTube, for videos. The interested reader can find the schemas, in case of structured and semi-structured sources, and the keywords, in case of unstructured sources, at the address `http://daisy.dii.univpm.it/dl/datasets/dl1`. The password to type is "`za.12&;lq74:#`".

It could appear that taking only 10 sources is excessively limited. However, we made this choice because we wanted to fully analyze the behavior and the performance of our approach and, as it will be clear, this requires the human intervention for verifying obtained results. This intervention would have become much more difficult with a higher number of sources to examine. At the same time, our test set is fully scalable. As a consequence, an interested reader, starting from the data sources provided at the address `http://daisy.dii.univpm.it/dl/datasets/dl1`, can construct a data set with a much higher number of sources, if necessary.

For our experiments, we used a server equipped with an Intel I7 Dual Core 5500U processor and 16 GB of RAM with the Ubuntu 16.04.3 operating system. Clearly, the capabilities of this server were limited. However, they were adequate for the (small) data set DS we have chosen to use in our tests.

Some important results we have found in this context are the following:

- We have defined an approach to create a structured representation of a natively unstructured data source.
- We have defined an approach to extract interschema properties and complex knowledge patterns from a data lake possibly consisting of a huge number of disparate data sources.

1.4.3 Internet of Things

Since the MIoT paradigm has been proposed for the first time by us in several papers connected with this thesis, there is no known case study or real example about it yet. As a consequence, in order to have a testbed for our experiments, we constructed a MIoT starting from some open data about things available on the Internet. In particular, we derived our data from *Thingful* [3]. This is a search engine for the Internet of Things, which allows us to search among a huge number of existing things, distributed all over the world. Thingful also provides some suitable APIs allowing the extraction of all the data we are looking for.

In order to construct our MIoT, we decided to work with 250 things whose data was derived from Thingful.

IoT	*Number of instances*
a.home	22
a.health	22
a.energy	22
a.transport	22
a.environment	22
b.near	14
b.mid	38
b.far	53
c.plain	44
c.hill	50
c.mountain	6

Table 1.1. Number of instances present in the IoTs of our MIoT

Our MIoT consists of 11 IoTs. We associated an object with each thing; therefore, we had 250 objects. In principle, for each object, we could have associated an instance for each dimension we aimed to investigate. However, in order to make our testbed closer to a generic MIoT, representing a real scenario, where it is not said that all the objects have exactly the same number of instances, we decided not to associate an instance with each dimension for all dimensions of our interest. Instead, we associated only one instance (distributed uniformly at random among the dimensions, and based on the features of the things of the IoTs of a given dimension) to 200 of the 250 objects. Analogously, we associated two instances (distributed by following the same guidelines mentioned above) to 35 of the 250 objects. Finally, we associated three instances to 15 of the 250 objects. At the end of this phase, we had 315 instances, distributed among the 11 IoTs of our MIoT, as shown in Table 1.1.

Some important results we have found in this context are the following:

- we have defined a new crawler, specifically conceived for our MIoT;
- we have defined a new approach to create topic-guided virtual IoTs.

1.4.4 Innovation Management

Data regarding patents adopted in our analyses has been taken from PATSTAT-ICRIOS database [108].

PATSTAT (i.e., EPO worldwide PATent STATistical database) is a database storing raw data about patents. It was constructed by EPO in cooperation with the World Intellectual Property Organization (WIPO), OECD and Eurostat. It is currently managed by EPO. It stores data about all patents, from 1978 to the current year, coming

from about 90 patent offices worldwide, comprising the most relevant ones, such as EPO and USPTO.

As pointed out above, data is registered in PATSTAT in a raw format. To facilitate its analysis, ICRIOS processed it and produced a cleaned and harmonized database, i.e., PATSTAT-ICRIOS. This includes all bibliographic variables concerning each patent application. In particular, it stores application number and date, publication number and date, priority, title and abstract, application status, designed states for protection, main and secondary International Patent Classification (IPC) codes, name and address of both the applicant and the inventor, references (i.e., citations) to prior-art patent and non-patent literature, the corresponding Nomenclature of Units for Territorial Statistics (NUTS3) and, finally, File Index concordance tables, allowing the conversion of IPC codes into more aggregated and manageable technological classes.

Some important results we have found in this context are the following:

- we have defined an approach to evaluate the scope of a patent;
- we have extracted knowledge regarding the lifecycle of a patent;
- we have defined new metrics specifically conceived to evaluate the innovation level of each country based on patent data.

1.5 Outline of this thesis

- In Part I, we will apply our network-based model and the associated social network-based approach to Biomedical Engineering, in particular to the investigation of neurological disorders. This part is organized as follows: in Chapter 2, we evaluate our approach on Creutzfeldt-Jacob Disease. In Chapter 3, we present results related to Mild Cognitive Impairment and Alzheimer's disease. In Chapter 4, we describe the application of our approach to Childhood Absence Epilepsy.
- In Part II, we apply our model and approach to Data Lakes. This part is organized as follows: In Chapter 5, we evaluate our approach to uniformly handle heterogeneous Data Lake sources. In Chapter 6, we present results related to inter-schema property derivation. Finally, in Chapter 7, we describe an approach to the extraction of complex knowledge patterns among concepts belonging to different sources.
- In Part III, we apply our model and approach to IoT. This part is organized as follows: in Chapter 8 we present its specialization to the extraction of knowledge from heterogeneous sensor data streams. In Chapter 9, we present the MIoT paradigm. Finally, in Chapter 10, we introduce the concept of topic-guided virtual IoTs in a MIoT.

- In Part IV, we apply our model and approach to Innovation Management. This part is organized as follows: in Chapter 11, we propose a well-tailored centrality measure for evaluating patents and their citations. In Chapter 12, we propose a new Social Network Analysis-based approach to extracting knowledge patterns about research activities and hubs in a set of countries. Finally, in Chapter 13, we introduce new metrics specifically conceived to evaluate the innovation level of each country based on patent data.
- Finally, in Part V, we draw some conclusions and delineate some possible future developments of our research efforts.

Part I

Neurological Disorders

In this part we apply our network-based model and the associated social network-based approach to help experts who investigate neurological disorders in which connections among brain areas play a key role. Our approach receives the EEG of a patient and associates a network with it, with nodes that represent electrodes and with edges that denote the disconnection degree of the corresponding brain areas. This part is organized as follows: in Chapter 2, we evaluate our approach on Creutzfeldt-Jacob Disease. In Chapter 3, we present results related to Mild Cognitive Impairment and Alzheimer's disease. In Chapter 4, we describe our approach applied to the evaluation of Childhood Absence Epilepsy.

2

Creutzfeldt Jakob Disease

2.1 Introduction

Creutzfeldt-Jacob Disease (CJD) is a rapidly progressive, uniformly fatal Transmissible Spongiform Encephalopathy (TSE). It is characterized by the accumulation of a variant of the host encoded cellular prion protein in the brain [463, 464, 282]. CJD became well known to common people some years ago because one of its variants, known as vCJD, has been linked to the transmission of the causative agent of the Bovine Spongiform Encephalopathy (BSE) to the human population, mainly in United Kingdom. Sporadic CJD (hereafter, sCJD) represents the most common form of CJD; in fact, it occurs worldwide in 84% of cases of CJD. It has an annual mortality rate of 1.39 per million.

An early and reliable diagnosis of CJD is extremely important to exclude other, potentially treatable, causes of rapidly progressive encephalopathies. However, the early diagnosis of this disease is complicated by the extreme heterogeneity of its clinical presentation [482, 140].

Electroencephalography (hereafter, EEG) has always been, and still is, one of the main methods to perform clinical diagnosis of neurological diseases in general [225, 337, 102, 369], and of CJD in particular. In fact, in the EEG of patients with sCJD, it is often possible to observe three-phase periodic spikes with sharp waves known as "Periodic Sharp Wave Complexes" (hereafter, PSWCs). More specifically, PSWCs were reported to occur in the EEG tracings of about two-thirds of patients with sCJD. For this reason, they were included in the World Health Organization diagnostic classification criteria of sCJD [463, 462, 464, 136].

In the past, approaches to investigating PSWCs in the EEGs of patients with sCJD were mainly based on signal processing [463, 421, 422, 464, 262, 17, 456, 420, 327, 223, 320]. By contrast, to the best of our knowledge, no network analysis based approach to investigating the CJD phenomenon has been previously proposed in the literature. Nevertheless, network analysis has been largely exploited in the investi-

gation of brain, especially in those application scenarios where brain connectivity is extremely important [385].

Since in sCJD (as well as in all the neurodegenerative diseases) the investigation of the connection level of the brain areas is extremely important, we argue that network analysis could play a key role in this research context. In particular, in the past literature, it was shown that, in presence of PSWCs, the areas of the brain are more connected than in absence of them [437]. This implies that network analysis could really represent a useful tool for investigating PSWCs.

In this chapter, we aim at providing a first contribution in this setting. Indeed, we propose a network analysis based approach to characterizing PSWCs in EEGs of patients with sCJD. Here, the term "characterizing" means two things, namely: *(i)* finding a quantitative coefficient - that we call *connection coefficient* - whose values are extremely different for the EEG tracing segments with PSWCs and the ones without PSWCs, and *(ii)* finding (possible) *network motifs* characterizing the presence (or, conversely, the absence) of PSWCs in an EEG tracing.

In our opinion, these two contributions are worthwhile. Indeed:

- It is true that PSWCs can be also seen with the naked eye by a human expert. However, the human eye can recognize PSWCs well only when these are marked, which happens in the advanced stages of the disease. By contrast, as previously pointed out, it could be extremely useful to carry out an early diagnosis of this disease, in which case PSWCs are not marked and cannot be recognized with the naked eye. A numeric coefficient can help to recognize PSWCs when they start to appear, thus allowing a much earlier diagnosis of sCJD. Furthermore, in the future, in presence of much more sophisticated electroencephalographs with 256 electrodes, the human eye could experiment much more difficulties in finding PSWCs.
- Differently from what generally happens in the network analysis literature (where motifs are intended as patterns occurring in a complex network much more frequently than they occur in randomized networks [312, 414, 339]), in our approach, motifs must be intended as network patterns occurring very frequently in the EEG tracing segments with PSWCs and very rarely in the ones without PSWCs, or vice versa. As a consequence, in our approach, motifs are found among different networks (and not in the same network). Our concept of motif has a twofold importance. First, motifs represent a further indicator of the presence of PSWCs. Second, and much more important, they could provide a characterization of the behavior of brain areas in presence of PSWCs. For instance, they could denote what are the brain areas most connected and/or most active in presence of PSWCs.

This is, probably, the most important contribution of our approach because this information cannot be directly derived by a human expert.

In network analysis, one of the most common and, at the same time, powerful tools to investigate the connection level of a network is the concept of *clique*. For this reason, our connection coefficient, as well as our definition of motifs and the corresponding motif extraction technique, are based on this concept, as will be clear in the following.

This chapter is organized as follows: in Section 2.2, we provide an overview of related literature. In Section 2.3, we present some support data structures. In Section 2.4, first we illustrate our connection coefficient, then we introduce our concept of motif and, finally, we present our approach to motif extraction.

2.2 Related Literature

In the last decades, PSWCs, along with very few other clinical criteria [462, 455, 475, 360, 167], have been recognized as capable of representing the most typical findings in the course of sCJD [463, 421, 422]. For instance, in [422], the authors investigate some issues concerning the diagnosis of CJD through PSWCs. Specifically, they measure sensitivity, specificity and the predictive values of PSWCs in cases where autopsy confirmed or excluded CJD. They find that PSWCs allowed a correct diagnosis in 64% of the CJD cases and returned false positives in 9% of other dementias. In [464], the authors study temporal and spatial development of EEG patterns in patients with sporadic or iatrogenic CJD. They show that Frontal Intermittent Rhythmical Delta Activity (hereafter, FIRDA) can be found in 4 out of 6 patients and, therefore, can be considered as an early EEG pattern associated with human prion diseases. They also show that FIRDA occurs at an early stage of CJD and is progressively replaced by PSWCs. In [406], the authors propose an approach to redefining several periodic patterns that can be found in the electroencephalograms of patients with sCJD. For this purpose, they exploit the criteria of the American Clinical Neurophysiology Society.

Most of the research about PSWCs focuses on their temporal evolution in the different stages of sCJD. In this context, [464, 262, 17, 45, 406] investigate: *(i)* the anomalies present in EEGs of patients with sCJD before the appearance of PSWCs, *(ii)* the fraction of patients with sCJD whose EEGs present PSWCs and, finally, *(iii)* the disappearance of PSWCs in several patients with sCJD close to death. For instance, in [262], the authors present a comparison between EEG findings in patients in the literature and 36 patients with CJD at the Massachusetts General Hospital. They found that 28 out of the 36 patients into examination had PSWCs at some time during their clinical course. PSWCs generally made their appearance within 12 weeks

of onset clinical symptoms. This result confirms what had already found in the past literature about this issue.

Other papers about this issue apply Independent Component Analysis (hereafter, ICA) to perform an early diagnosis of sCJD through the analysis of PSWCs. For instance, in [456], the authors use ICA to split typical PSWCs into several independent temporal components, in conjunction with spatial maps. They also show that ICA may increase the sensitivity of EEG and facilitate the early diagnosis of CJD. More recent approaches aim at performing an early diagnosis of sCJD through Support Vector Machine (hereafter, SVM) or Deep Learning. Specifically, in [387], the authors use the principal component analysis to reduce the dimensionality of the dataset. Then, they exploit an SVM-based algorithm to analyze and classify EEG signals. In [320], a technique to distinguish the EEG of patients with early-stage CJD from other forms of rapidly progressive dementia is proposed. This technique reaches an average accuracy of 89%, an average sensitivity of 92%, and an average specificity of 89%.

Further papers deepen the investigations of EEGs with PSWCs by means of non linear analysis [420] and, more in general, the relationship between the acquisition of measures composing an EEG (possibly preceded by a pre-processing step [185, 58]) and the possible presence of anomalies in the corresponding tracing. For instance, in [420], the authors analyze the EEG of a patient with CJD using the method of non linear forecasting. They aim at re-examining the hypothesis that PSWCs reflect non linear, possibly chaotic, dynamics of the cortical networks. They show that these episodes can be predicted much better than the irregular background activity. As a consequence, they prove the usefulness of non linear models to gain a better understanding of brain dynamics. These models show that oscillations are an intrinsic property of the system and external noise only modifies them to a some extent.

To the best of our knowledge, only two papers tried to determine the most active brain areas in presence of PSWCs [327, 223]. Specifically, in [327], the authors apply computerized topographic analysis to study periodic discharges (of which PSWCs represent an example) in EEGs. They investigate how the presence of periodic diphasic or triphasic sharp wave discharges evolves as long as the disease evolves over time. They also investigate the local distribution of these discharges and find that even those patterns, which seemed generalized on a visual inspection, are not truly bilaterally symmetrical nor synchronous. Even in the same patient, separate discharges have different focal onset areas and reach peak maximum activity in diverse brain areas. In [223], the authors show that Dipole Source Localization (hereafter, DSL) can provide information about the source location of particular EEG activities. They also show that PSWCs in CJD are generalized discharges and may have multiple cortical sources or alternating activation pathways in cortical areas. Furthermore, they combine ICA

and DSL to define sources of generalized discharges, such as PSWCs. To characterize the dipole sources responsible for PSWCs across patients into evaluation, they perform cluster analysis using k-means. This way, they find a few sets of dipole clusters and try to explain the pathophysiological mechanisms of PSWCs based on these results.

Finally, an interesting property of PSWCs is investigated in [437]. Here, the authors show that, in patients with CJD, fusions of neuronal processes, in particular dendrites, may lead to abnormal electrotonic coupling between cells, which causes powerful excitatory interaction whereby large neuronal aggregates burst in near synchrony. This cortical synchronous discharges would give rise to PSWCs in the electroencephalogram; in the meantime, similar discharges in brainstem, spinal cord, or elsewhere could lead to myoclonic jerks.

As pointed out in the Introduction, to the best of our knowledge, our paper represents the first attempt to apply network analysis to investigate PSWCs in patients with sCJD. Nevertheless, network analysis has been frequently applied in the investigation of modern brain mapping techniques, such as diffusion MRI, EEG and MEG [385]. As a matter of fact, brain networks are complex and may hence be characterized by applying complex network analysis methods.

Network analysis allows brain networks to be reliably quantified by means of a small number of neurobiologically meaningful and easily computable measures [415, 15, 57, 197, 187]. Furthermore, comparisons of structural or functional network topologies between subject population can reveal possible connectivity abnormalities in neurological and psychiatric disorders [419, 56, 259, 370, 454].

Measures of individual network elements typically quantify connectivity profiles and, hence, reflect how these elements are embedded in the network. Example of these measures are: *(i) functional segregation*, i.e., the ability of specialized processing to occur within densely interconnected groups of brain regions [459, 330]; *(ii) functional integration*, i.e., the ability to rapidly combine specialized information from distributed brain regions [14]; *(iii) paths in functional networks*, i.e., sequences of distinct nodes and links representing potential routes of information flow between pairs of brain regions [206, 459, 253]; *(iv) anatomical motifs*, i.e., patterns of local connectivity in a given network whose significance is determined by their frequency within that network [312, 414, 339].

Since anatomical brain connectivity influences the neuropathological lesions' capability of affecting functional brain activity, network analysis can be exploited to characterize the resilience of the brain networks to these lesions. Indirect measures of resilience quantify anatomical features reflecting network vulnerability to insults. Two examples of indirect measures are degree distribution [52] and assortativity coefficient [329]. Direct measures of network resilience generally test the network before

and after a possible insult. The effects of lesions on the network may be quantified by characterizing the changes in the resulting anatomical connectivity, or in the emergent simulated functional connectivity or dynamical activity [31]. An approach to investigating EEGs through complex networks was recently proposed in [132]. Here, the authors construct suitable weighted complex networks and apply community structure detection techniques to them for analyzing multi-channel EEG signals. Then, they show that this method is well suited in identifying epileptic seizures in EEGs.

2.3 Basic Support Data Structures

The EEGs to perform our investigation were provided by three different Italian centers (i.e., University "Magna Graecia" of Catanzaro, Neurologic Institute "Carlo Besta" of Milano, and Neurologic Institute of the University of Catania). They regard a group of ten patients with sCJD examined in the last 15 years in these three centers[1]. The EEGs were recorded through scalp electrodes placed according to the international 10-20 system. The specific montage was: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz and Pz, where G2 (located between electrodes Fz and Cz) was used as reference. The EEG were recorded in the morning in a comfortable, eye closed, resting state. The mean recording length is 20 min. The EEG was high-pass filtered at 0.5 Hz, low-pass filtered at 70 Hz, and a 50 Hz notch filter was also applied. The EEG traces were then downsampled to 256 Hz. The artifactual epochs were visually detected and marked by the EEG experts and later excluded from the analysis. Cleaned data was stored in a MongoDB database [6]. To give an idea of it, we report some of its features: *(i)* size = 357.8 MB; *(ii)* number of collections = 10 ; *(iii)* number of objects = 20; *(iv)* number of indexes = 10; *(v)* index size = 160 KB.

We segmented each EEG at disposal in such a way as to separate the tracing segments with PSWCs from those without PSWCs (Figure 2.1). As a consequence, for each EEG, we had several tracing segments, which could be grouped in two distinct sets, namely, those containing PSWCs and those not containing PSWCs.

Formally speaking, let $EEGSet$ be the set of EEGs at our disposal, let eeg be an EEG of $EEGSet$. Starting from eeg, it is possible to define a network $\mathcal{N}$ (resp., $\overline{\mathcal{N}}$) representing the set of segments of eeg with PSWCs (resp., without PSWCs). Specifically:

$$\mathcal{N} = \langle V, E \rangle \qquad\qquad \overline{\mathcal{N}} = \langle V, \overline{E} \rangle$$

[1] We are aware that the number of patients under examination is low. However, this is due to the fact that sCJD is a very rare disease and, consequently, it is very difficult to collect data about it.

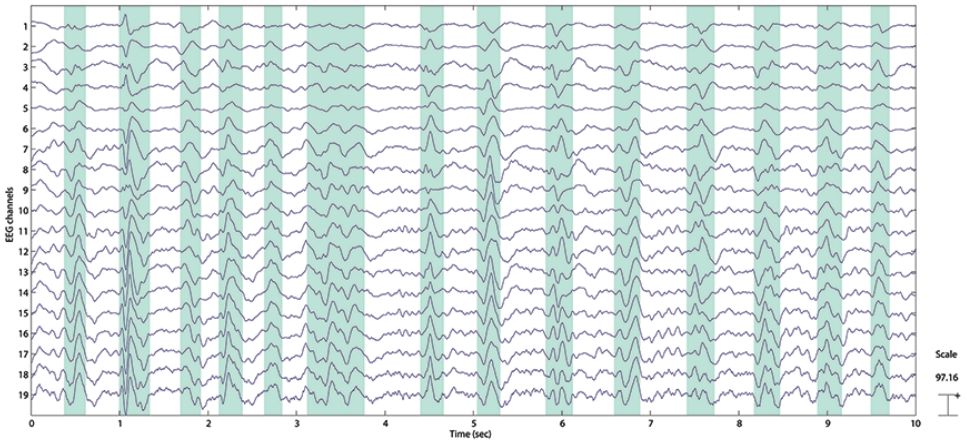


Fig. 2.1. Partitioning of an EEG into segments with PSWCs and without PSWCs - shaded segments correspond to the ones with PSWCs

Here, V is the set of the nodes of $\mathcal{N}$ and $\overline{\mathcal{N}}$. Each node $v_i \in V$ corresponds to an electrode. Since, in our EEGs, the electrodes were applied by following the 10-20 system, we have that $|V| = 19$.

E (resp., $\overline{E}$) is the set of the edges of $\mathcal{N}$ (resp., $\overline{\mathcal{N}}$). Each edge $e_{ij} \in E$ connects the nodes v_i and v_j. It can be represented as:

$$e_{ij} = (v_i, v_j, w_{ij})$$

Here, w_{ij} is a measure of "distance" between v_i and v_j. This "distance" is an indicator of the disconnection level of v_i and v_j. Actually, each measure representing this characteristic could be adopted in our model. In the experiments described in this chapter, we used the *Permutation Join Entropy* (PJE) between v_i and v_j, which is a new metric of cross-randomness between channels in multivariate electrophysiological time-series [112, 293].

In order to make our model more "user-friendly" and "expressive" and, at the same time, more capable of discriminating strong and weak connections between brain areas, we decided to construct two new networks, namely $\mathcal{N}_\pi$ and $\overline{\mathcal{N}_\pi}$, obtained from $\mathcal{N}$ and $\overline{\mathcal{N}}$ by removing the edges with an "excessive" weight[2] and by coloring the other ones on the basis of their weight. More specifically, blue edges denote strong connections (i.e., small weights), red edges represent intermediate ones and, finally, green edges indicate weak connections.

Formally speaking, let $EEGSet$ be the set of EEGs at our disposal and let $NSet$ (resp., $\overline{NSet}$) be the set of the corresponding networks. As usual, $\mathcal{N} \in NSet$ (resp., $\overline{\mathcal{N}} \in \overline{NSet}$) corresponds to an EEG of $EEGSet$ and denotes its tracing segments with

[2] Recall that, in our model, edge weight is a measure of distance.

(resp., without) PSWCs. Let $max(\cdot)$ be a function returning the maximum weight of the edges of the network provided in input to it. Then, it is possible to define:

$$Max = Med_{\mathcal{N} \in NSet}(max(\mathcal{N})) \qquad \overline{Max} = Med_{\overline{\mathcal{N}} \in \overline{NSet}}(max(\overline{\mathcal{N}}))$$

In other words, Max (resp., $\overline{Max}$) represents the median of the maximum weights of the edges of the networks of $NSet$ (resp., $\overline{NSet}$). The choice of the *Median* function is motivated by the exigency to make our approach robust against possible outliers or noise.

On the basis of these two parameters, we can define $\mathcal{N}_\pi$ (resp., $\overline{\mathcal{N}_\pi}$), obtained from $\mathcal{N}$ (resp., $\overline{\mathcal{N}}$) by removing the edges with the highest weights and by coloring the other ones according to the rules mentioned above. Specifically:

$$\mathcal{N}_\pi = \langle V, E_\pi \rangle \qquad \overline{\mathcal{N}_\pi} = \langle V, \overline{E_\pi} \rangle$$

Here, the nodes of $\mathcal{N}_\pi$ and $\overline{\mathcal{N}_\pi}$ are the same as the ones of $\mathcal{N}$ and $\overline{\mathcal{N}}$.

In order to define the sets E_π and $\overline{E_\pi}$, first we compute the value $min(Max, \overline{Max})$ and, then, we divide the interval $[0, min(Max, \overline{Max})]$ in 10 equiwidth intervals. We indicate with $len = \frac{min(Max, \overline{Max})}{10}$ the length of each of these intervals. We denote with $\mathcal{I}_k$ the k^{th} interval; it includes the values of the weights higher than or equal to $(k-1) \cdot len$ and lesser than or equal to $k \cdot len$.

Now, we are able to specify the structure of E_π and $\overline{E_\pi}$:

$$E_\pi = \{(v_i, v_j, w_{ij}) \in E \mid w_{ij} \notin \mathcal{I}_k, 8 \leq k \leq 10\}$$
$$\overline{E_\pi} = \left\{(v_i, v_j, w_{ij}) \in \overline{E} \mid w_{ij} \notin \mathcal{I}_k, 8 \leq k \leq 10\right\}$$

In these formulas, the values of k have been determined experimentally. The factor $min(Max, \overline{Max})$ allows a more selective filtering in such a way as to lower the effects of possible noise or outliers. Observe that the same threshold is used in the definition of both E_π and $\overline{E_\pi}$. In fact, we need the same "severity" level of filtering to better characterize the possible differences and peculiarities of $\mathcal{N}_\pi$ against $\overline{\mathcal{N}_\pi}$, and vice versa.

Now, after having defined E_π (resp., $\overline{E_\pi}$), we can "color" the edges composing it. Specifically:

$$E_\pi = E_\pi^b \cup E_\pi^r \cup E_\pi^g \qquad \overline{E_\pi} = \overline{E_\pi^b} \cup \overline{E_\pi^r} \cup \overline{E_\pi^g}$$

where:

$$E_\pi^b = \{(v_i, v_j, w_{ij}) \in E_\pi \mid w_{ij} \in \mathcal{I}_k, k < 5\}$$
$$E_\pi^r = \{(v_i, v_j, w_{ij}) \in E_\pi \mid (w_{ij} \in \mathcal{I}_k, k = 5\}$$
$$E_\pi^g = \{(v_i, v_j, w_{ij}) \in E_\pi \mid w_{ij} \in \mathcal{I}_k, k > 5\}$$

In an analogous way, it is possible to define $\overline{E_\pi^b}$, $\overline{E_\pi^r}$ and $\overline{E_\pi^g}$. In Figure 2.2, we report the networks $\mathcal{N}$ and $\overline{\mathcal{N}}$ for a patient with sCJD. The disposal of the nodes in

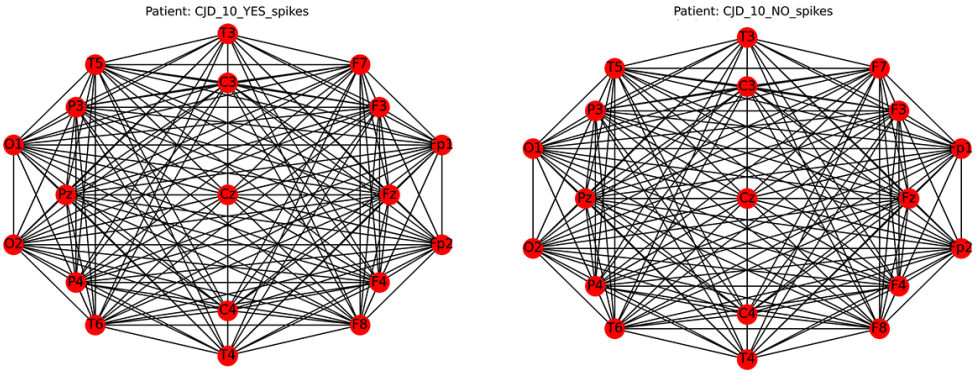


Fig. 2.2. Original Networks $\mathcal{N}_\pi$ and $\overline{\mathcal{N}_\pi}$ for the patient CJD 10

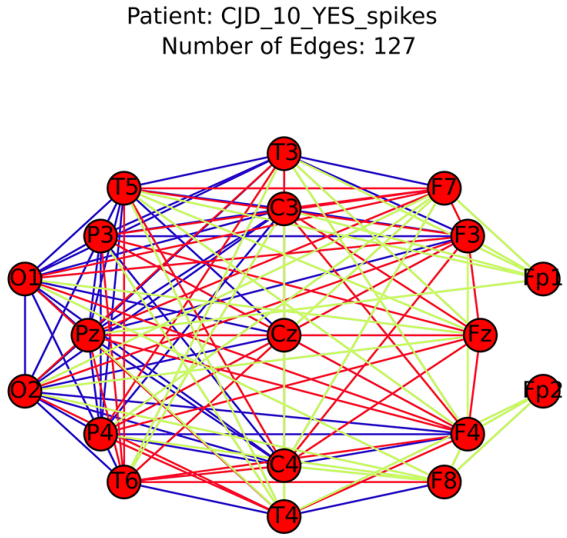


Fig. 2.3. Colored Network $\mathcal{N}_\pi$ for the patient CJD 10

the networks reflects the 10-20 system, even if they are rotated 90 degrees clockwise. From the analysis of this figure, it is possible to observe that the two networks are indistinguishable. Indeed, the only difference would be in the edge weights, which are not reported due to layout reasons. Instead, in Figures 2.3 and 2.4, we illustrate the corresponding colored networks $\mathcal{N}_\pi$ and $\overline{\mathcal{N}_\pi}$. Observe how the filtering of the edges with the highest distance, along with the coloration of the other ones on the basis of the closeness of the corresponding nodes, make this model very expressive.

The trends emerging from these figures have been confirmed in all the other EEGs at our disposal. In particular, we observe that:

- For a specific EEG, $\mathcal{N}_\pi$ has more edges than $\overline{\mathcal{N}_\pi}$; in fact, the average number of edges of $\mathcal{N}_\pi$ is 142, whereas the one of $\overline{\mathcal{N}_\pi}$ is 134. Furthermore, the edges of $\mathcal{N}_\pi$

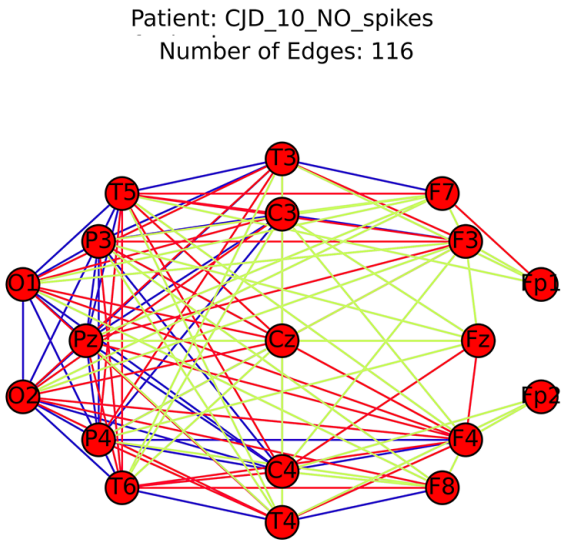


Fig. 2.4. Colored Network $\overline{\mathcal{N}_\pi}$ for the patient CJD 10

are generally stronger than the ones of $\overline{\mathcal{N}_\pi}$; indeed, the average weight of the edges of $\mathcal{N}_\pi$ is 1.66, whereas the one of the edges of $\overline{\mathcal{N}_\pi}$ is 1.92[3].

- In both networks, the strongest edges can be found in the occipital area of the skull.

2.4 PSWC Characterization

This section represents the "core" of our paper because it aims at providing two network analysis based ways to characterize PSWCs. The former is a numeric coefficient, whereas the latter consists of graph-based knowledge patterns.

2.4.1 Connection Coefficient

As pointed out in the Introduction, one of the main features to investigate in neurodegenerative patients is the connection level of brain areas. This feature is also relevant in the problem we are addressing. In fact, as pointed out in Section 2.2, in the literature, it was shown that, in presence of PSWCs, brain areas are more connected than in absence of them [437]. Furthermore, in Section 2.3, we have seen that the networks corresponding to the tracing segments with PSWCs are generally more numerous and more strongly connected than the networks corresponding to the tracing segments without PSWCs.

[3] Recall that edge weights are distance measures.

In network analysis, one of the most important (and, at the same time, simple and basic) tools for investigating network connection is the concept of clique. We recall that, given a network, a clique of dimension k represents a totally connected subnetwork with k nodes. On the basis of this reasoning, a quantitative coefficient for discriminating the networks corresponding to the tracing segments with PSWCs from the ones associated with the tracing segments without PSWCs could highly benefit from cliques.

In particular, this coefficient should take the following considerations into account:

- Both the dimension and the number of cliques are important as connectivity indicators.
- The concept of clique is intrinsically exponential; in other words, a clique of dimension $n+1$ is exponentially more complex than a clique of dimension n.
- It is necessary to avoid the possible presence of outliers and noise; as a consequence, it is inappropriate to consider only the cliques with the maximum dimension; by contrast, it appears more equilibrate to consider also the cliques with the maximum, sub-maximum and sub-sub-maximum dimension. On the other hand, it appears unnecessary and time consuming to consider the other cliques because their contribution decreases exponentially against their dimension.

Starting from these considerations, we now define our connection coefficient. Let $\mathcal{N}_\pi = \langle V, E_\pi \rangle$ be a colored network representing the tracing segments with PSWCs of an EEG of a patient. Let $\mathcal{C}$ be the set of cliques of $\mathcal{N}_\pi$ and let $dim(\cdot)$ be a function returning the dimension of a set of cliques, all of the same dimension, received in input. Then, it is possible to define:

- the subset $\mathcal{C}_{M_1} \subset \mathcal{C}$ of the cliques with the maximum dimension;
- the subset $\mathcal{C}_{M_2} \subset \mathcal{C}$ of the cliques with the sub-maximum dimension;
- the subset $\mathcal{C}_{M_3} \subset \mathcal{C}$ of the cliques with the sub-sub-maximum dimension.

In the same way, $\overline{\mathcal{C}_{M_1}}$, $\overline{\mathcal{C}_{M_2}}$ and $\overline{\mathcal{C}_{M_3}}$ can be defined.

Finally, let $|\mathcal{C}_{M_1}|$, $|\mathcal{C}_{M_2}|$ and $|\mathcal{C}_{M_3}|$ be the cardinalities (i.e., the number of cliques) of $\mathcal{C}_{M_1}$, $\mathcal{C}_{M_2}$ and $\mathcal{C}_{M_3}$, respectively.

Then, the connection coefficient $cc_{\mathcal{N}_\pi}$, associated with $\mathcal{N}_\pi$, is defined as:

$$cc_{\mathcal{N}_\pi} = \sum_{i=1}^{3} \left(\mid \mathcal{C}_{M_i} \mid \cdot 2^{dim(\mathcal{C}_{M_i})} \right)$$

This formula takes all the above considerations into account. In an analogous way, it is possible to define the connection coefficient $cc_{\overline{\mathcal{N}_\pi}}$ associated with $\overline{\mathcal{N}_\pi}$.

In Table 2.1 (resp., 2.2) we report the values of $|\mathcal{C}_{M_i}|$ (resp., $|\overline{\mathcal{C}_{M_i}}|$) and $dim(\mathcal{C}_{M_i})$ (resp., $dim(\overline{\mathcal{C}_{M_i}})$), $1 \leq i \leq 3$, as well as the values of $cc_{\mathcal{N}_\pi}$ (resp., $cc_{\overline{\mathcal{N}_\pi}}$), for all the patients at our disposal. Finally, in Table 2.3, we report the values of $cc_{\mathcal{N}_\pi}$ and $cc_{\overline{\mathcal{N}_\pi}}$,

along with the percentage of decrease observed when passing from $cc_{\mathcal{N}_\pi}$ to $cc_{\overline{\mathcal{N}_\pi}}$, for the same patients.

Patient	$dim(\mathcal{C}_{M_1})$	$\lvert\mathcal{C}_{M_1}\rvert$	$dim(\mathcal{C}_{M_2})$	$\lvert\mathcal{C}_{M_2}\rvert$	$dim(\mathcal{C}_{M_3})$	$\lvert\mathcal{C}_{M_3}\rvert$	$cc_{\mathcal{N}_\pi}$
CJD 02	14	4	9	1	4	1	66064
CJD 04	12	5	8	1	7	1	20864
CJD 05	15	1	14	2	12	1	69632
CJD 08	16	3	6	1	0	0	196672
CJD 09	14	2	13	2	9	1	49664
CJD 10	13	1	12	3	8	1	20736
CJD 13	12	2	11	1	10	2	12288
CJD 16	11	1	10	6	8	6	9721
CJD 19	19	1	0	0	0	0	524288
CJD 22	16	2	14	2	9	1	164352

Table 2.1. Values of $dim(\mathcal{C}_{M_i})$, $|\mathcal{C}_{M_i}|$ $(1 \leq i \leq 3)$ and $cc_{\mathcal{N}_\pi}$ for all the patients at our disposal

Patient	$dim(\overline{\mathcal{C}_{M_1}})$	$\lvert\overline{\mathcal{C}_{M_1}}\rvert$	$dim(\overline{\mathcal{C}_{M_2}})$	$\lvert\overline{\mathcal{C}_{M_2}}\rvert$	$dim(\overline{\mathcal{C}_{M_3}})$	$\lvert\overline{\mathcal{C}_{M_3}}\rvert$	$cc_{\overline{\mathcal{N}_\pi}}$
CJD 02	12	1	11	10	6	1	24640
CJD 04	12	2	11	2	10	1	13312
CJD 05	14	2	13	1	10	2	43008
CJD 08	13	6	11	1	9	1	51712
CJD 09	12	6	10	1	8	1	25856
CJD 10	11	4	9	2	8	2	9728
CJD 13	10	4	9	2	8	1	5376
CJD 16	11	2	10	4	9	2	9216
CJD 19	18	2	0	0	0	0	524288
CJD 22	14	4	12	2	8	1	78080

Table 2.2. Values of $dim(\overline{\mathcal{C}_{M_i}})$, $|\overline{\mathcal{C}_{M_i}}|$ $(1 \leq i \leq 3)$ and $cc_{\overline{\mathcal{N}_\pi}}$ for all the patients at our disposal

From the analysis of these tables we can draw two important results. In fact:

- $cc_{\mathcal{N}_\pi}$ is always higher than $cc_{\overline{\mathcal{N}_\pi}}$ except for the patient CJD 19 for whom the two coefficients have the same value. However, $\mathcal{N}_{\pi_{19}}$ and $\overline{\mathcal{N}_{\pi_{19}}}$ are associated with a very particular EEG. As an evidence of this fact, we observe that $\mathcal{N}_{\pi_{19}}$ is totally connected and, therefore, has only a unique clique coinciding with it. $\overline{\mathcal{N}_{\pi_{19}}}$, instead, is totally connected except for only one edge; as a consequence, it has only two cliques, each consisting of 18 nodes.

  As a consequence, we can say that connection coefficient is really a quantitative parameter capable of characterizing the tracing segments with PSWCs.

Patient	$cc_{\mathcal{N}_\pi}$	$cc_{\overline{\mathcal{N}_\pi}}$	$\frac{cc_{\overline{\mathcal{N}_\pi}} - cc_{\mathcal{N}_\pi}}{cc_{\mathcal{N}_\pi}}$
CJD 02	66064	24640	-62.70%
CJD 04	20864	13312	-36.20%
CJD 05	69632	43008	-38.24%
CJD 08	196672	51712	-73.71%
CJD 09	49664	25856	-47.94%
CJD 10	20736	9728	-53.09%
CJD 13	12288	5376	-56.25%
CJD 16	9721	9216	-5.19%
CJD 19	524288	524288	0.00%
CJD 22	164352	78080	-52.49%

Table 2.3. Values of $cc_{\mathcal{N}_\pi}$, $cc_{\overline{\mathcal{N}_\pi}}$ and $\frac{cc_{\overline{\mathcal{N}_\pi}} - cc_{\mathcal{N}_\pi}}{cc_{\mathcal{N}_\pi}}$ for all the patients at our disposal

- The values obtained for $cc_{\mathcal{N}_\pi}$ and $cc_{\overline{\mathcal{N}_\pi}}$ confirm the previous results presented in the literature about the fact that brain areas are more connected to each other in presence of PSWCs than in absence of them [437].

2.4.2 Motifs

In the previous section, we have seen that connection coefficient, strongly based on cliques, is capable of characterizing the tracing segments with PSWCs. In this section, we aim at investigating the possible presence of motifs characterizing the tracing segments with PSWCs against the ones without PSWCs, and vice versa.

Actually, motifs have been already investigated and exploited in past approaches adopting network analysis (see, for instance, [312, 414, 339]). In those scenarios, they are considered as [312]:

> "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks".

In our approach, we use motifs in a completely different fashion. Indeed, we do not examine a unique complex network to find patterns frequently repeated therein. By contrast, we search for patterns appearing frequently in the networks corresponding to the tracing segments with PSWCs and being absent in the networks corresponding to the tracing segments without PSWCs, thus characterizing the former segment typology against the latter, and vice versa.

As will be clear in the following, our approach to deriving motifs exploits the support data structures introduced in Section 2.3, along with a further support network, strongly based on the clique concept, which we call clique network. We describe this data structure in the next subsection.

Clique network

Let eeg be an EEG of $EEGSet$, let $\mathcal{N}_\pi$ (resp., $\overline{\mathcal{N}_\pi}$) be the colored network associated with the tracing segments with PSWCs (resp., without PSWCs) of eeg and let $\mathcal{C}$ (resp., $\overline{\mathcal{C}}$) be the set of cliques of $\mathcal{N}_\pi$ (resp., $\overline{\mathcal{N}_\pi}$).

The clique network $\mathcal{CN}$ (resp., $\overline{\mathcal{CN}}$), corresponding to $\mathcal{N}_\pi$ (resp., $\overline{\mathcal{N}_\pi}$) and $\mathcal{C}$ (resp., $\overline{\mathcal{C}}$), is defined as:

$$\mathcal{CN} = \langle CV, CE \rangle \qquad\qquad \overline{\mathcal{CN}} = \langle \overline{CV}, \overline{CE} \rangle$$

Here:

- CV represents the set of the nodes of $\mathcal{CN}$. There is a node $v_i \in CV$ for each node $v_i \in V$. A weight w_i is associated with v_i; it represents the number of cliques of $\mathcal{C}$ in which v_i is involved.
- CE indicates the set of the edges of $\mathcal{CN}$. There is an edge $(v_i, v_j, w_{ij}) \in CE$ if the edge (v_i, v_j) is present in at least one clique of $\mathcal{C}$. w_{ij} denotes the number of cliques of $\mathcal{C}$ in which (v_i, v_j) is present.
- $\overline{CV}$ and $\overline{CE}$ are analogous to CV and CE, but for $\overline{\mathcal{C}}$, instead of for $\mathcal{C}$.

The edges of $\mathcal{CN}$ can be "colored" in a way analogous to the edges of $\mathcal{N}_\pi$. Also in this case, blue edges are the strongest ones, red edges have an intermediate strongness and green edges are the weakest ones. Formally speaking:

$$CE = CE^b \cup CE^r \cup CE^g$$

- $CE^b = \{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE,\ w_{ij} > th^{rb}\}$
- $CE^r = \{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE,\ (w_{ij} > th^{gr}) \wedge (w_{ij} \leq th^{rb})\}$
- $CE^g = \{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE, w_{ij} \leq th^{gr}\}$

Also in this case, we determined th^{rb} and th^{gr} experimentally. In particular, we found that the best values for them are $th^{rb} = 6$ and $th^{gr} = 3$. In an analogous fashion, we defined $\overline{CE}$.

In Figures 2.5 and 2.6, we report the clique networks $\mathcal{CN}$ and $\overline{\mathcal{CN}}$ associated with the patient CJD 16. Here, the dimension of nodes is directly proportional to their weights. In these figures, there are two graphical indicators that help the reader to understand the features of the tracing segments with PSWCs and the ones without PSWCs. In fact, the color of an edge (which, we recall, is directly connected to the corresponding weight) is an indicator of the strongness of the connection between the corresponding brain areas. The dimension of a node (directly connected to the associated weight) is an indicator of the connection degree of the brain area associated with it and, ultimately, an indicator of its activity level.

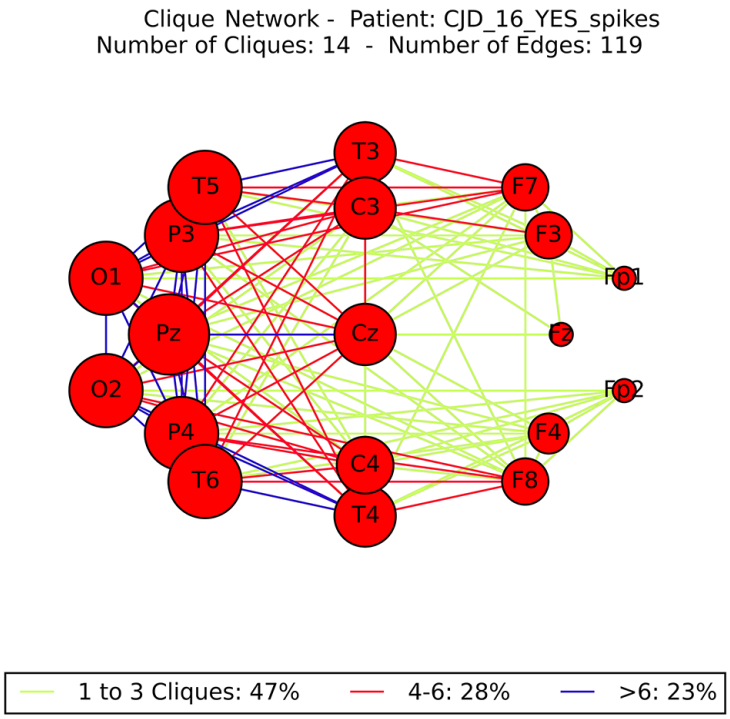


Fig. 2.5. Clique Network $\mathcal{CN}$ for the patient CJD 16

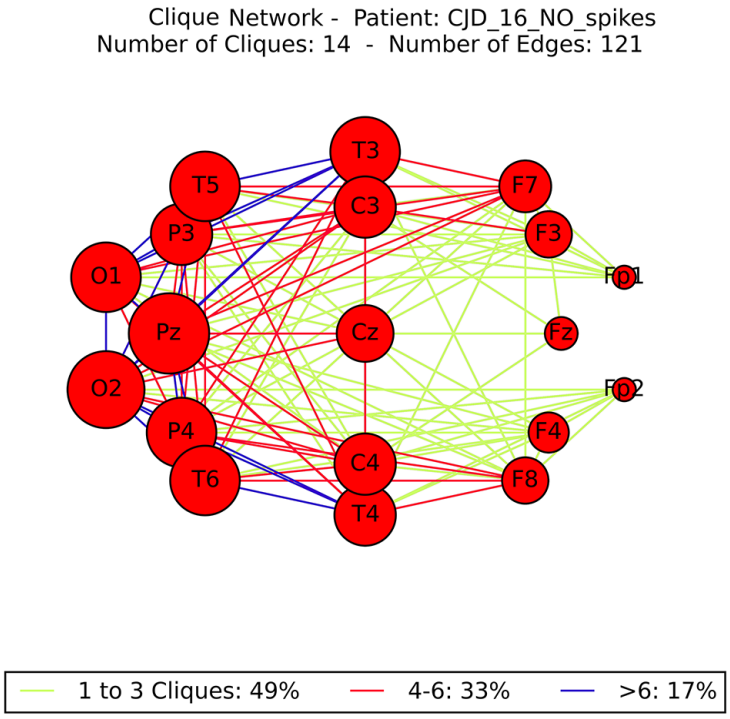


Fig. 2.6. Clique Network $\overline{\mathcal{CN}}$ for the patient CJD 16

Clique-based Support Data Structures and Parameters

After having introduced clique social networks, we are now able to fully present the data structures and parameters adopted in our motif extraction approach. Specifically,

let eeg be an EEG of $EEGSet$, let $\mathcal{N}_\pi$ and $\overline{\mathcal{N}_\pi}$ be the corresponding colored networks and let $\mathcal{CN}$ and $\overline{\mathcal{CN}}$ be the corresponding clique networks. We recall that:

$$\mathcal{N}_\pi = \langle V, E_\pi \rangle \qquad \overline{\mathcal{N}_\pi} = \langle V, \overline{E_\pi} \rangle \qquad \mathcal{CN} = \langle CV, CE \rangle \qquad \overline{\mathcal{CN}} = \langle \overline{CV}, \overline{CE} \rangle$$

We also recall that:

$$E_\pi = E_\pi^b \cup E_\pi^r \cup E_\pi^g \quad \overline{E_\pi} = \overline{E_\pi^b} \cup \overline{E_\pi^r} \cup \overline{E_\pi^g} \quad CE = CE^b \cup CE^r \cup CE^g$$
$$\overline{CE} = \overline{CE^b} \cup \overline{CE^r} \cup \overline{CE^g}$$

Now, we define more restrictive colored networks and clique networks by removing green edges from the networks defined previously. Specifically, we define:

$$\mathcal{N}_{\pi\pi} = \langle V, E_{\pi\pi} \rangle \qquad \overline{\mathcal{N}_{\pi\pi}} = \langle V, \overline{E_{\pi\pi}} \rangle \qquad \mathcal{CN}_{\pi\pi} = \langle CV, CE_{\pi\pi} \rangle$$
$$\overline{\mathcal{CN}_{\pi\pi}} = \langle \overline{CV}, \overline{CE_{\pi\pi}} \rangle$$

where:

$$E_{\pi\pi} = E_{\pi\pi}^b \cup E_{\pi\pi}^r \quad \overline{E_{\pi\pi}} = \overline{E_{\pi\pi}^b} \cup \overline{E_{\pi\pi}^r} \quad CE_{\pi\pi} = CE_{\pi\pi}^b \cup CE_{\pi\pi}^r$$
$$\overline{CE_{\pi\pi}} = \overline{CE_{\pi\pi}^b} \cup \overline{CE_{\pi\pi}^r}$$

After this, we introduce the sets $\mathcal{C}_\pi$, $\overline{\mathcal{C}_\pi}$, $\mathcal{C}_{\pi\pi}$, $\overline{\mathcal{C}_{\pi\pi}}$, $\mathcal{NC}_{\pi\pi}$, $\overline{\mathcal{NC}_{\pi\pi}}$ as the sets of the cliques of $NSet_\pi$, $\overline{NSet_\pi}$, $NSet_{\pi\pi}$, $\overline{NSet_{\pi\pi}}$, $CNSet_{\pi\pi}$, $\overline{CNSet_{\pi\pi}}$, respectively.

Starting from these last sets, we define the sets $\mathcal{T}_\pi$, $\overline{\mathcal{T}_\pi}$, $\mathcal{T}_{\pi\pi}$, $\overline{\mathcal{T}_{\pi\pi}}$, $\mathcal{TC}_{\pi\pi}$, $\overline{\mathcal{TC}_{\pi\pi}}$ as the sets of totally connected triads of $\mathcal{C}_\pi$, $\overline{\mathcal{C}_\pi}$, $\mathcal{C}_{\pi\pi}$, $\overline{\mathcal{C}_{\pi\pi}}$, $\mathcal{NC}_{\pi\pi}$, $\overline{\mathcal{NC}_{\pi\pi}}$[4]. Then, we define the sets $NSet_\pi$ (resp., $\overline{NSet_\pi}$, $NSet_{\pi\pi}$, $\overline{NSet_{\pi\pi}}$, $CNSet_{\pi\pi}$, $\overline{CNSet_{\pi\pi}}$) comprising all the networks $\mathcal{N}_\pi$ (resp., $\overline{\mathcal{N}_\pi}$, $\mathcal{N}_{\pi\pi}$, $\overline{\mathcal{N}_{\pi\pi}}$, $\mathcal{CN}_{\pi\pi}$, $\overline{\mathcal{CN}_{\pi\pi}}$) associated with the EEGs of $EEGSet$.

Finally, let t be a generic triad. We call $nocc_\pi$ (resp., $\overline{nocc_\pi}$, $nocc_{\pi\pi}$, $\overline{nocc_{\pi\pi}}$, $cnocc_{\pi\pi}$, $\overline{cnocc_{\pi\pi}}$) the number of occurrences of t in $NSet_\pi$ (resp., $\overline{NSet_\pi}$, $NSet_{\pi\pi}$, $\overline{NSet_{\pi\pi}}$, $CNSet_{\pi\pi}$, $\overline{CNSet_{\pi\pi}}$).

Extraction of basic motifs

After having defined all support data structures and parameters, we are able to describe our motif extraction approach. It consists of two main steps, the former devoted to the extraction of basic motifs and the latter conceived to the construction of derived ones. In this section, we focus on the former, whereas, in the next section, we present the latter. Preliminarily, it is necessary to specify what is a basic motif in our context. Specifically:

[4] We recall that a triad is a subnetwork consisting of three nodes. The totally connected triad is considered the most stable structure in network analysis. Clearly, a totally connected triad can be considered as a clique of dimension 3.

Let t be a totally connected triad of $NSet_\pi$. If (1) t is present *frequently* in $NSet_\pi$ and is *absent* in $\overline{NSet_\pi}$, and if (2) this trend is confirmed (also only to a lesser extent) for $NSet_{\pi\pi}$ and $\overline{NSet_{\pi\pi}}$ and also for $CNSet_{\pi\pi}$ and $\overline{CNSet_{\pi\pi}}$, then t is a basic motif. In particular, t is a motif characterizing the tracing segments with PSWCs against the ones without PSWCs.

In order to be able to really extract basic motifs, it is necessary to provide a quantitative definition of this rule. For this purpose, it is preliminarily necessary to associate a numeric value with the concept of "frequently". Recalling that all the sets $NSet_\pi$, $\overline{NSet_\pi}$, $NSet_{\pi\pi}$, $\overline{NSet_{\pi\pi}}$, $CNSet_{\pi\pi}$, $\overline{CNSet_{\pi\pi}}$ have the same cardinality, we can define the following threshold:

$$th_f = \alpha_f \cdot |NSet_\pi|$$

A high value of α_f would lead to a high value of th_f and, therefore, to a low number of basic motifs. In this case, the correctness of results is privileged over their completeness and our approach becomes restrictive. Vice versa, a low value of α_f would cause a low value of th_f and, therefore, would produce a high number of basic motifs. In this case, the completeness of results is privileged over their correctness and the approach becomes permissive. We experimentally set the value of α_f to 0.30, which we chose as the default one of our approach. In fact, this value experimentally proved to be the most "equilibrate" (i.e., neither extremely permissive nor extremely restrictive) one.

Therefore, let $t \in \mathcal{T}_\pi$ be a totally connected triad of $NSet_\pi$. If, with reference to t, the following conditions simultaneously hold:

- (1) $(nocc_\pi \geq th_f) \wedge (\overline{nocc_\pi} = 0)$
- (2) $(nocc_{\pi\pi} > \overline{nocc_{\pi\pi}}) \wedge (cnocc_{\pi\pi} > \overline{cnocc_{\pi\pi}})$

then, t is a basic motif characterizing the tracing segments with PSWCs against the ones without PSWCs.

In a dual fashion, it is possible to define the basic motifs associated with $\overline{NSet_\pi}$ and characterizing the tracing segments without PSWCs against the ones with PSWCs.

In the following, we indicate by $\mathcal{M}_\pi$ (resp., $\overline{\mathcal{M}_\pi}$) the set of motifs extracted starting from the triads of $NSet_\pi$ (resp., $\overline{NSet_\pi}$).

In an analogous way, it is possible to derive the basic motifs of the sets $\mathcal{M}_{\pi\pi}$, $\overline{\mathcal{M}_{\pi\pi}}$, $\mathcal{CM}_{\pi\pi}$ and $\overline{\mathcal{CM}_{\pi\pi}}$, obtained starting from the triads of $NSet_{\pi\pi}$, $\overline{NSet_{\pi\pi}}$, $CNSet_{\pi\pi}$ and $\overline{CNSet_{\pi\pi}}$.

Figure 2.7 represents two basic motifs belonging to $\mathcal{CM}_{\pi\pi}$ and $\overline{\mathcal{CM}_{\pi\pi}}$, obtained by applying our approach to the data at our disposal.

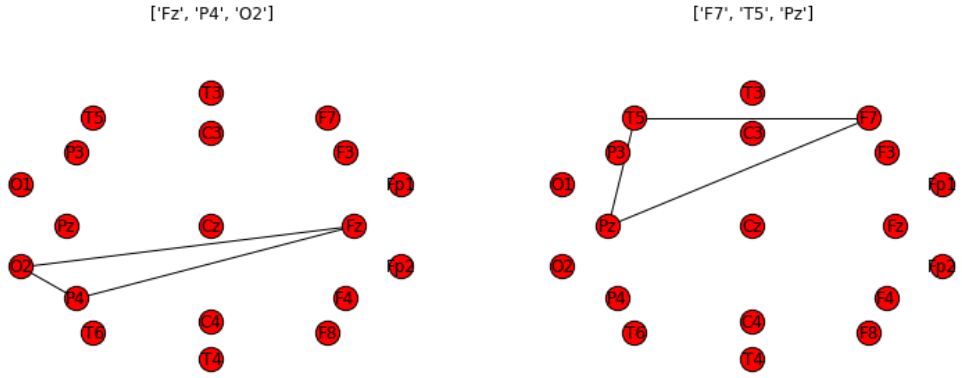


Fig. 2.7. Two basic motifs belonging to $\mathcal{CM}_{\pi\pi}$ (at left) and $\overline{\mathcal{CM}_{\pi\pi}}$ (at right)

Observe that a motif is not only an indicator of the tracing segments with PSWCs (or without PSWCs). Actually, it is much more. Indeed, it allows a characterization of the behavior of the brain areas in presence (resp., in absence) of PSWCs. For instance, it denotes what are the brain areas most connected in presence (resp., in absence) of PSWCs. This is, probably, the most important contribution of our approach because this information cannot be directly derived by a human expert. The basic motifs derived by our approach, with α_f set to its default value of 0.30, are reported in Table 2.4.

$\mathcal{M}_\pi$	$\overline{\mathcal{M}_\pi}$	$\mathcal{M}_{\pi\pi}$	$\overline{\mathcal{M}_{\pi\pi}}$	$\mathcal{CM}_{\pi\pi}$	$\overline{\mathcal{CM}_{\pi\pi}}$
				['Cz', 'Fz', 'P4']	['F4', 'P4', 'T6']
				['Fz', 'P4', 'Pz']	['F4', 'O2', 'T6']
				['Fz', 'O2', 'P4']	['C4', 'F4', 'T6']
				['Cz', 'Fz', 'Pz']	['F7', 'P3', 'Pz']
				['Cz', 'Fz', 'O2']	['F7', 'O2', 'P3']
				['Fz', 'O2', 'Pz']	['P4', 'T3', 'T4']
					['F7', 'Pz', 'T5']
					['F7', 'Pz', 'T3']
					['F7', 'O2', 'Pz']
					['F7', 'O1', 'Pz']
					['F7', 'O2', 'T5']
					['F7', 'O2', 'T3']
					['F7', 'O1', 'O2']
					['Pz', 'T3', 'T4']
					['T3', 'T4', 'T5']
					['T3', 'T4', 'T6']
					['O2', 'T3', 'T4']
					['O1', 'T3', 'T4']

Table 2.4. The basic motifs extracted by our approach with α_f set to its default value of 0.30

$\mathcal{M}_\pi$	$\overline{\mathcal{M}_\pi}$	$\mathcal{M}_{\pi\pi}$	$\overline{\mathcal{M}_{\pi\pi}}$	$\mathcal{CM}_{\pi\pi}$	$\overline{\mathcal{CM}_{\pi\pi}}$
['F8', 'Fz', 'P3']		['C3', 'Fz', 'O1']		['Fz', 'P4', 'T6']	['Cz', 'F4', 'P4']
['Fz', 'T3', 'T4']		['C3', 'Fz', 'O2']		['Fz', 'O2', 'P4']	['F4', 'P4', 'T4']
['F8', 'Fz', 'O1']		['Fz', 'O1', 'P3']		['Cz', 'Fz', 'P4']	['C4', 'F4', 'P4']
['F4', 'F8', 'Fz']		['Fz', 'O2', 'P3']		['Fz', 'P3', 'P4']	['F4', 'O2', 'P4']
['Cz', 'F8', 'Fz']		['Cz', 'Fz', 'O1']		['Fz', 'O1', 'P4']	['F4', 'P4', 'T6']
['C4', 'F8', 'Fz']		['C4', 'Fz', 'O1']		['Fz', 'P4', 'Pz']	['Cz', 'F4', 'O2']
['F8', 'Fz', 'T4']		['Fz', 'O1', 'Pz']		['Fz', 'O2', 'T6']	['Cz', 'F4', 'T4']
['F8', 'Fz', 'Pz']		['Fz', 'O1', 'P4']		['Cz', 'Fz', 'T6']	['C4', 'Cz', 'F4']
['F8', 'Fz', 'P4']		['Fz', 'O1', 'O2']		['Fz', 'P3', 'T6']	['Cz', 'F4', 'T6']
['F8', 'Fz', 'T6']		['F4', 'Fz', 'O2']		['Fz', 'O1', 'T6']	['F4', 'O2', 'T4']
['F8', 'Fz', 'O2']		['Cz', 'Fz', 'O2']		['Fz', 'Pz', 'T6']	['C4', 'F4', 'O2']
['F4', 'Fz', 'T3']		['C4', 'Fz', 'O2']		['Cz', 'Fz', 'O2']	['F4', 'O2', 'T6']
['F3', 'F4', 'Fp2']		['Fz', 'O2', 'Pz']		['Fz', 'O2', 'P3']	['C4', 'F4', 'T4']
['F4', 'Fp2', 'T5']		['Fz', 'O2', 'P4']		['Fz', 'O1', 'O2']	['F4', 'T4', 'T6']
		['F3', 'P4', 'T3']		['Fz', 'O2', 'Pz']	['C4', 'F4', 'T6']
		['C4', 'Fp2', 'P4']		['Cz', 'Fz', 'P3']	['F4', 'P4', 'Pz']
		['Fp2', 'P4', 'T6']		['Cz', 'Fz', 'O1']	['F4', 'O2', 'Pz']
		['F8', 'Fp2', 'P4']		['Cz', 'Fz', 'Pz']	['C4', 'F4', 'Pz']
		['F7', 'T3', 'T4']		['Fz', 'O1', 'P3']	['F4', 'Pz', 'T6']
		['F8', 'T3', 'T4']		['Fz', 'P3', 'Pz']	['F7', 'Pz', 'T3']
		['C3', 'T3', 'T4']		['Fz', 'O1', 'Pz']	['F7', 'T3', 'T6']
		['Cz', 'T3', 'T4']			['F7', 'O2', 'T3']
		['C4', 'T3', 'T4']			['F7', 'Pz', 'T5']
		['T3', 'T4', 'T5']			['F7', 'T5', 'T6']
		['P3', 'T3', 'T4']			['F7', 'O2', 'T5']
		['Pz', 'T3', 'T4']			['F7', 'O1', 'Pz']
		['O2', 'T3', 'T4']			['F7', 'O1', 'T6']
		['T3', 'T4', 'T6']			['F7', 'O1', 'O2']
		['O1', 'T3', 'T4']			['F7', 'Pz', 'T6']
					['F7', 'P3', 'Pz']
					['F7', 'O2', 'Pz']
					['F7', 'P3', 'T6']
					['F7', 'O2', 'T6']
					['F7', 'O2', 'P3']
					['C3', 'T5', 'T6']
					['C3', 'C4', 'T6']
					['T3', 'T4', 'T5']
					['O1', 'T3', 'T4']
					['Pz', 'T3', 'T4']
					['Cz', 'T3', 'T6']
					['T3', 'T4', 'T6']
					['C4', 'F4', 'P3']
					['O2', 'T3', 'T4']
					['Cz', 'P4', 'T3']
					['P4', 'T3', 'T4']
					['F4', 'P3', 'P4']

Table 2.5. The basic motifs extracted by our approach with α_f set to 0.20

From the analysis of this table, it emerges that all the basic motifs extracted by our approach belong to $\mathcal{CM}_{\pi\pi}$ and $\overline{\mathcal{CM}_{\pi\pi}}$. However, in the columns of Table 2.4, we have reported all the six possible sets to evidence that, if the human expert wants to be more "permissive", she can decrease the value of α_f w.r.t. the default one. In this case, she could find basic motifs also in the other sets. Just to give an idea of this last case, in Table 2.5, we report the basic motifs extracted by our approach with α_f set to 0.20. Observe that: *(i)* much more basic motifs have been found; *(ii)* obtained motifs belong not only to $\mathcal{CM}_{\pi\pi}$ and $\overline{\mathcal{CM}_{\pi\pi}}$ but also to $\mathcal{M}_\pi$ and $\mathcal{M}_{\pi\pi}$.

On the other hand, we cannot exclude that, in presence of sets of available EEGs richer than the one at our disposal, some basic motifs could appear in each possible set also when α_f is set to its default value of 0.30.

Extraction of derived motifs

Once basic motifs have been extracted, and a first version of $\mathcal{M}_\pi$, $\overline{\mathcal{M}_\pi}$, $\mathcal{M}_{\pi\pi}$, $\overline{\mathcal{M}_{\pi\pi}}$, $\mathcal{CM}_{\pi\pi}$ and $\overline{\mathcal{CM}_{\pi\pi}}$ has been obtained, it is possible to construct derived (and, possibly, much more complex and significant) motifs starting from them.

Our approach constructs new derived motifs starting from the already known ones. It uses nodes common to two or more known motifs as "junction points". Formally speaking, let $m_1 = \langle V_1, E_1 \rangle$ and $m_2 = \langle V_2, E_2 \rangle$ be two motifs of $\mathcal{M}_\pi$ such that $V_1 \cap V_2 \neq \emptyset$. Then, it is possible to construct a candidate motif as the union of m_1 and m_2:

$$m_{12} = \langle V_1 \cup V_2, E_1 \cup E_2 \rangle$$

Once m_{12} has been constructed, analogously to what we have seen for basic motifs, it is necessary to evaluate $nocc_\pi$, $\overline{nocc_\pi}$, $nocc_{\pi\pi}$, $\overline{nocc_{\pi\pi}}$, $cnocc_{\pi\pi}$ and $\overline{cnocc_{\pi\pi}}$[5]. If, for these parameters, conditions (1) and (2) presented in Section 2.4.2 hold, then m_{12} can be added to $\mathcal{M}_\pi$, i.e., $\mathcal{M}_\pi = \mathcal{M}_\pi \cup \{m_{12}\}$.

Clearly, the addition of a new motif in $\mathcal{M}_\pi$ could lead to the possibility that new candidate motifs are constructed. As a consequence, the enrichment process of $\mathcal{M}_\pi$ is iterative and terminates when, during an iteration, no new motif is added to $\mathcal{M}_\pi$. In an analogous fashion, the derived motifs of $\overline{\mathcal{M}_\pi}$, $\mathcal{M}_{\pi\pi}$, $\overline{\mathcal{M}_{\pi\pi}}$, $\mathcal{CM}_{\pi\pi}$ and $\overline{\mathcal{CM}_{\pi\pi}}$ can be extracted.

In Figures 2.8, 2.9 and 2.10 we report the most significant derived motifs extracted by our approach. The motif in Figure 2.8 derives from the tracing segments with PSWCs. It indicates that, in presence of PSWCs, the most active areas of the human brain reside in its right part. The motifs shown in Figures 2.9 and 2.10 derive from the tracing segments without PSWCs. They indicate that, in absence of PSWCs, the most active areas of the human brain reside in its left part (Figure 2.9) and in its occipital part (Figure 2.10).

Clearly, these results will require much more efforts and investigations in the future, especially by experts in neurological diseases, in order to understand their complete meaningfulness. Nevertheless, they are an interesting "food for thought" that our approach is providing to researchers in this sector.

Once all possible derived motifs have been obtained, it is possible to construct two sets, namely:

$$\mathcal{GM} = \mathcal{M}_\pi \cup \mathcal{M}_{\pi\pi} \cup \mathcal{CM}_{\pi\pi} \qquad \overline{\mathcal{GM}} = \overline{\mathcal{M}_\pi} \cup \overline{\mathcal{M}_{\pi\pi}} \cup \overline{\mathcal{CM}_{\pi\pi}}$$

[5] Clearly, for derived motifs, $nocc_\pi$, $\overline{nocc_\pi}$, $nocc_{\pi\pi}$, $\overline{nocc_{\pi\pi}}$, $cnocc_{\pi\pi}$ and $\overline{cnocc_{\pi\pi}}$ refer to the number of occurrences on motifs, instead of on triads.

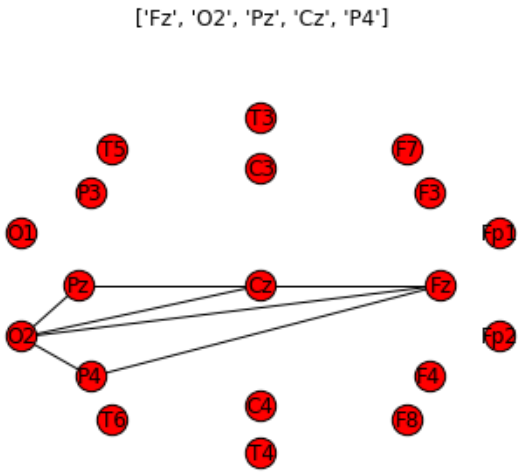


Fig. 2.8. The most significant motif characterizing the tracing segments with PSWCs

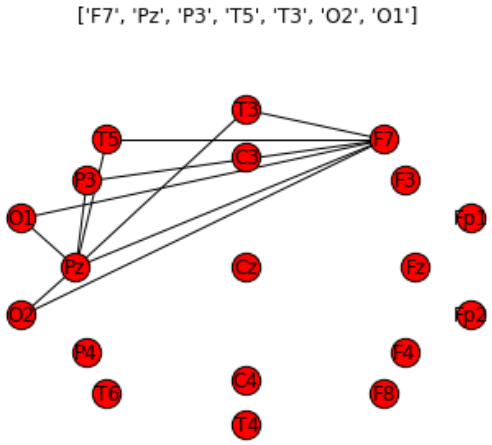


Fig. 2.9. One of the most significant motifs characterizing the tracing segments without PSWCs

The former contains all motifs corresponding to the tracing segments with PSWCs, whereas the latter comprises all motifs representing the tracing segments without PSWCs.

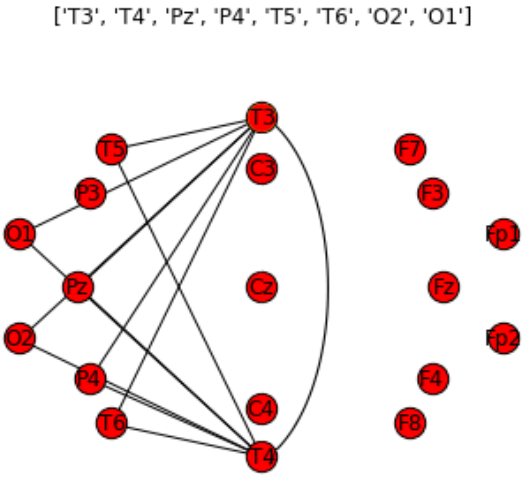


Fig. 2.10. A further significant motif characterizing the tracing segments without PSWCs

3

Mild Cognitive Impairment - Alzheimer's disease (AD)

3.1 Introduction

3.1.1 Motivations and Related Literature

In recent years, the incidence of Alzheimer's Disease (hereafter, AD) is growing because the population is aging in most countries. For this reason, the efforts to design approaches capable of determining the onset of this disease in advance are intensifying [198, 374]. Even if this issue is challenging, it is extremely complex, as also evidenced in past literature. As a matter of fact, it was shown that: *(i)* AD shares many clinical features with other forms of dementia, and *(ii)* the molecular pathomechanism of AD becomes active several years before neurons start dying and cognitive deficits appear. For a definitive diagnosis of AD, the biopsy of brain tissues is even necessary.

A further important issue that makes the diagnosis on these patients difficult concerns the fact that they, just by the very nature of their disease, do not easily undergo examinations, like Magnetic Resonance Imaging, which force them to stay motionless for a long time.

A non-invasive and well tolerated examination, which can be done on patients with neurological disorders, is ElectroEncephaloGram (hereafter, EEG) [392, 231]. Indeed, in scientific literature, several signal theory-based approaches employing EEG to investigate patients with Mild Cognitive Impairment (hereafter, MCI) or AD have been proposed [117, 119, 250, 322, 321].

At the same time, taking into account that an EEG can be easily modeled as a network, with nodes that represent electrodes and edges that denote connections between electrodes, several approaches to investigating neurodegenerative diseases have been recently proposed [385, 122, 169].

As specified in [364], MCI can be prodromal for AD. In fact, several papers suggest that patients with MCI tend to convert to AD with a rate of about 10-15% annually [120]. For this reason, a large variety of approaches aiming at characterizing both MCI

and AD have been proposed in past literature. Several of these approaches are based on the analysis of EEG.

The possibility of using EEG for characterizing patients with AD is evidenced in [220, 207]. In fact, the EEGs of patients with AD present some peculiarities, namely slowing, reduced complexity and perturbations in synchrony. However, it was shown that these effects can be observed with different intensities in different patients. For this reason, at present, none of them alone allows a reliable diagnosis of AD at an early stage.

Several approaches investigate the slowing of EEGs in patients with AD (see, for instance, [115, 444]). In particular, some of these papers also investigate the effect of AD in the tracings of EEGs in the sub-bands α, β, δ and θ. The changes in spectral power are determined by means of Fourier Transform [115, 444] or sparsified time-frequency maps [444]. Other approaches analyze the reduced complexity of EEG signals in patients with AD (see, for instance, [207, 63]). In this context, to quantify this reduction, the authors apply several measures, namely approximate entropy [207], auto mutual information [207], sample entropy [207], multiscale entropy [207], Lempel-Ziv complexity [207], and fractal dimension [63, 355]. Finally, further approaches investigate the decrease of synchrony in patients with MCI and AD w.r.t. age-matched control subjects (see, for instance, [115, 119]). To quantify this decrease, many measures have been proposed, e.g., Pearson correlation coefficient [119], coherence [119, 395], Granges causality [119], information-theoretic [119], state space-based synchrony measures [115, 119], phase synchrony indices [115, 119] and stochastic event synchrony [119].

Few studies evidence an increase of EEG synchrony in patients, recorded during working memory task [468]. This inverse effect is often interpreted as the result of a compensatory mechanism in the brain. Several papers (e.g., [18, 34]) examine the changes of brain activity in patients with MCI using MagnetoEncephaloGram (MEG), instead of EEG.

Network analysis [21, 127, 69, 179] has been frequently applied in the investigation of modern brain mapping techniques. Indeed, it provides several neurobiologically meaningful and easily computable measures [197, 187] to reliably quantify the main characteristics of brain networks. Furthermore, it is extremely useful to detect possible connectivity abnormalities characterizing neurological and psychiatric disorders [370, 454]. Typical network analysis parameters and structures adopted for this purpose are functional segregation [459, 330], functional integration [14], paths in functional networks [206], anatomical motifs [312, 414, 339]. Network analysis was also adopted to quantify the resilience of brain to insults [31].

Several papers (e.g., [122, 472, 442, 218]) focus on the usage of network analysis to investigate MCI or AD through the EEGs of the corresponding patients (an overview of these studies can be found in [315]). The parameter generally adopted to measure the connection level of brain areas is clustering coefficient, even if other basic network analysis parameters, such as characteristic path length, global efficiency, connectivity degree and connectivity density, have been proven able to partially evidence the loss of connectivity characterizing the progression of AD [319]. In some cases, these measures are applied not only to the overall EEG but also to one or more sub-bands (for instance, [417] considers the β sub-band). In [240], the authors investigate the spatial distribution of EEG phase synchrony in patients with AD. For this purpose, they analyze the surface topography of the Multivariate Phase Synchronization of multichannel EEG. They investigate these features for both the overall EEG and its sub-bands.

3.1.2 Objectives and general description of the proposed approach

This chapter presents a network analysis-based approach to help experts in their analyses of subjects with MCI and AD and their evolution over time. The inputs of our approach are the EEGs of the patients to analyze, performed at time t_0 and, then again three months later, at time t_1.

Given an EEG of a patient, our approach constructs a network with nodes that represent the electrodes and edges that denote connections between electrodes. Each edge has associated a weight representing a measure of the connection level between the brain areas covered by the corresponding electrodes.

Once the network associated with an EEG has been constructed, it is possible to employ the enormous wealth of knowledge already existing in network analysis to face the issues of our interest. In particular, since it is well known that, in AD progression and in MCI progression towards AD, a key role is played by the loss of connectivity among the different cortical areas, it appears reasonable to start our analysis from the knowledge on connectivity gained in network analysis in the past. Here, one of the most important tools available for this purpose is the concept of clique. We recall that a clique of dimension k in a network represents a completely connected subnetwork formed by k nodes.

Our approach applies the concept of clique to construct a suitable data structure, which we call *clique network*, and an indicator of the connectivity level of the brain areas, called *connection coefficient*, allowing us to distinguish patients with MCI from patients with AD. This indicator or, better, a second one constructed starting from it and called *conversion coefficient*, which associates the quantification of connection loss with the probability that such a loss corresponds to MCI conversion to AD, has

proven particularly useful in helping experts to understand if a patient with MCI is converting to AD. In our opinion, connection and conversion coefficients represent a first relevant contribution of our paper. Indeed, the literature lacks longitudinal studies on MCI/AD, due to the difficulty in keeping such patients and their caregivers loyal to a periodical follow-up program. We believe that the present research can be a starting point for motivating other people to engage longitudinal studies on MCI and AD.

We have striven to, at least partially, face the issue of the availability of a limited-size database by performing a further experimental campaign on virtual patients with MCI or AD, suitably constructed from the real ones (see Section 3.3.1). Furthermore, our approach might be extended on other neurological disorders, related to an impairment of cortical connectivity (Parkinson's disease [427] [202], schizophrenia [105, 22], epilepsy [242, 445], ADHD [19] and autism [20]).

In addition, our approach aims at facing a second issue. In fact, it aims at verifying if *network motifs* exist, i.e., specific sub-networks, or network patterns, which are very frequent in one kind of patient and absent, or very rare, in the other. Also for this issue we have obtained interesting results, since we have found some motifs characterizing patients with MCI from patients with AD. Interestingly, our concept of motif has a further, much more important, feature. Indeed, it could provide a characterization of the behavior of brain areas in presence of a disorder (or when a patient converts from a disorder to another). For instance, motifs could denote what brain areas are more connected and/or more active in presence of MCI and in absence of AD or, dually speaking, what brain areas are most affected or damaged when a patient with MCI converts to AD. As for this topic, the results obtained by our approach are very similar to the ones obtained by the approach described in [240], acquired by applying a completely different methodology.

Besides these two major contributions, this chapter presents some minor ones. For instance, our analysis confirms the previous results, obtained in past literature through completely different approaches [457, 160, 118, 64], about the capability of helping experts to understand if a patient with MCI converts to AD, which characterizes the tracings of some of the four sub-bands (i.e., α, β, δ and θ) of an EEG. In particular, according to past results obtained in the literature, we have shown that the sub-bands δ and θ play a key role in this context. Furthermore, we introduce the connection coefficient. This parameter is strictly dependent on both the number and the dimension of the cliques that can be found in the network. Since cliques represent completely connected subnetworks, connection coefficient is well suited as an indicator of the connection degree of a network. Actually, as we will show below, connection coefficient shows a much better performance than clustering coefficient, which is the

parameter classically adopted in Social Network Analysis to measure the connectivity degree of a network.

This chapter is organized as follows: in Section 3.2, we illustrate the proposed approach in detail. In Section 3.3, we describe the results of the experimental campaign we conducted to determine the adequacy of our approach and discuss them.

3.2 Methods

3.2.1 Input and Support Data Structures

The input of our approach consists of a set $EEGSet$ of EEGs at our disposal. It has the following structure:

$$EEGSet = \{CtrlSet, MCISet_0, ADSet_0, MCISet_1, ADSet_1\}$$

where: *(i)* $CtrlSet$ is the set of the EEGs of the control subjects; *(ii)* $MCISet_0$ (resp., $MCISet_1$) is the set of the EEGs of the patients with MCI at t_0 (resp., t_1); *(iii)* $ADSet_0$ (resp., $ADSet_1$) is the set of the EEGs of the patients with AD at t_0 (resp., t_1).

Starting, Colored and Clique Networks

Let eeg be an EEG[1] of $EEGSet$. Starting from eeg, it is possible to define a network:

$$\mathcal{N} = \langle V, E \rangle$$

Here, V is the set of nodes of $\mathcal{N}$. Each node $v_i \in V$ corresponds to an electrode of the EEG. In our EEGs, electrodes were applied by following the 10-20 system and $|V| = 19$.

E is the set of the edges of $\mathcal{N}$. Each edge e_{ij} connects the nodes v_i and v_j and can be represented as:

$$e_{ij} = (v_i, v_j, w_{ij})$$

Here, w_{ij} is a measure of "distance" between v_i and v_j. It is an indicator of the disconnection level of v_i and v_j. Even if our approach is orthogonal to the measure adopted for estimating synchrony, in our experiments we chose to employ PDI (*Permutation Disalignment Index*), which proved to be well suited in quantifying the overall coupling strength between EEG signals associated with MCI progression towards AD [292]. In particular, PDI was compared with Coherence and Dissimilarity

[1] At this moment, we do not make any assumptions about the subject whom *eeg* refers to. She/he could be a control subject, a patient with MCI or a patient with *AD*.

Index, a nonlinear and symbolic measure that proved to be promising in the pairwise analysis of EEG data. PDI was shown to outperform both Coherence and Dissimilarity Index [292]. It can help whenever a multivariate, amplitude invariant, robust to noise, nonlinear coupling strength analysis is necessary. All the above mentioned features are useful in EEG processing because EEG is multivariate, influenced by the distance from the reference electrode, affected by noise and nonlinear behavior. For all these reasons, in our experiments, w_{ij} was set to the average PDI between v_i and v_j. The interested reader can find a detailed description of PDI in Appendix A.1.

In order to make our model more "user-friendly" and "expressive" and, at the same time, more capable of discriminating strong and weak connections between the different brain areas, we decided to construct a new network, namely $\mathcal{N}_\pi$, obtained from $\mathcal{N}$ by removing the edges with an "excessive" weight (see below) and by coloring the others on the basis of their weight. As a matter of fact, edges with an "excessive" weight represent connections between portions of the brain having a low connection degree. In particular, blue edges denote strong connections (i.e., small weights), red edges represent intermediate ones and, finally, green edges indicate weak connections. In the following, we formalize this reasoning:

$$\mathcal{N}_\pi = \langle V, E_\pi \rangle$$

Here, the nodes of $\mathcal{N}_\pi$ are the same as the ones of $\mathcal{N}$. To define E_π, we employ the distribution of the weights of the edges of $\mathcal{N}$. Specifically, let max_E (resp., min_E) be the maximum (resp., minimum) weight of an edge of E. Starting from them, it is possible to define a parameter $step_E = \frac{max_E - min_E}{10}$, which represents the length of a "step" of the interval between min_E and max_E. We can define $d^k(E)$, $0 \leq k \leq 9$, as the number of the edges of E with weights that belong to the interval between $min_E + k \cdot step_E$ and $min_E + (k+1) \cdot step_E$. All these intervals are closed on the left and open on the right, except for the last one that is closed both on the left and on the right. We are now able to formalize E_π. Specifically, it consists of all the edges of E belonging to $d^k(E)$, where $k \leq th_{max}$.

Now, we can "color" the edges composing E_π. Specifically, $E_\pi = E_\pi^b \cup E_\pi^r \cup E_\pi^g$. Here:

- $E_\pi^b = \left\{ e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{min} \leq k \leq th_{br}} d^k(E) \right\}$;
- $E_\pi^r = \left\{ e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{br} < k \leq th_{rg}} d^k(E) \right\}$;
- $E_\pi^g = \left\{ e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{rg} < k \leq th_{max}} d^k(E) \right\}$.

In this definition, we determined the bounds of E_π^b, E_π^r and E_π^g experimentally. In particular, we set the values of th_{min}, th_{br}, th_{rg} and th_{max} to 0, 1, 4 and 6, respectively. From this definition, it is clear that discarded edges are those belonging to the eighth, ninth and tenth intervals of the range $[min_E, max_E]$.

To give an idea of the expressiveness of colored networks, in Figure 3.1 we report the distribution of the edge weights and the colored network of a control subject (resp., a patient with MCI, a patient with AD). The disposal of nodes in the network reflects the 10-20 system, even if they are rotated 90 degrees clockwise. It is straightforward to observe that the control subject presents a weight distribution more biased on the left than the patient with MCI, who, in turn, presents a weight distribution more biased on the left than the patient with AD. A direct consequence of this fact is that the colored network of the patient with AD presents lesser and weaker edges than the colored network of the patient with MCI that, in turn, presents lesser and weaker edges than the colored network of the control subject.

In order to quantify this phenomenon, in Table 3.1 we report the values of some measures characterizing the three colored networks shown in the three figures above. Specifically, the considered measures are: *(i)* the total number of colored edges; *(ii)* the total number of blue (resp., red, green) edges[2]; *(iii)* the percentage of colored edges against the total number of original edges; *(iv)* the percentage of blue (resp., red, green) edges against the total number of original edges. The quantitative results reported in Table 3.1, fully confirm the qualitative analysis mentioned above.

Parameter	*Control Subject*	*Patient with MCI*	*Patient with AD*
Total number of colored edges	170	141	69
Total number of blue edges	105	35	2
Total number of red edges	59	75	40
Total number of green edges	6	31	27
Percentage of colored edges	99.4%	82.5%	40.3%
Percentage of blue edges	61.4%	20.5%	1.2%
Percentage of red edges	34.5%	43.8%	23.4%
Percentage of green edges	3.5%	18.1%	15.8%

Table 3.1. Quantitative results representing the networks of Figure 3.1

As pointed out in the Introduction, the concept of clique[3] can play a key role in the investigation of those neurological diseases, like MCI and AD, where it is extremely important to analyze the connection level between brain areas. For this reason, in our approach, we introduce a further support data structure, called clique network.

In particular, let eeg be an EEG of $EEGSet$, let $\mathcal{N}_\pi = \langle V, E_\pi \rangle$ be the corresponding colored network and let $\mathcal{C}$ be the set of the cliques of $\mathcal{N}_\pi$. The clique network $\mathcal{CN}$, corresponding to $\mathcal{N}_\pi$ and $\mathcal{C}$, is defined as:

$$\mathcal{CN} = \langle CV, CE \rangle$$

Here:

[2] Recall that blue edges are the strongest ones, red edges have an intermediate weight, whereas green edges are the weakest ones.

[3] Recall that a clique of dimension k in a network represents a completely connected subnetwork formed by k nodes.

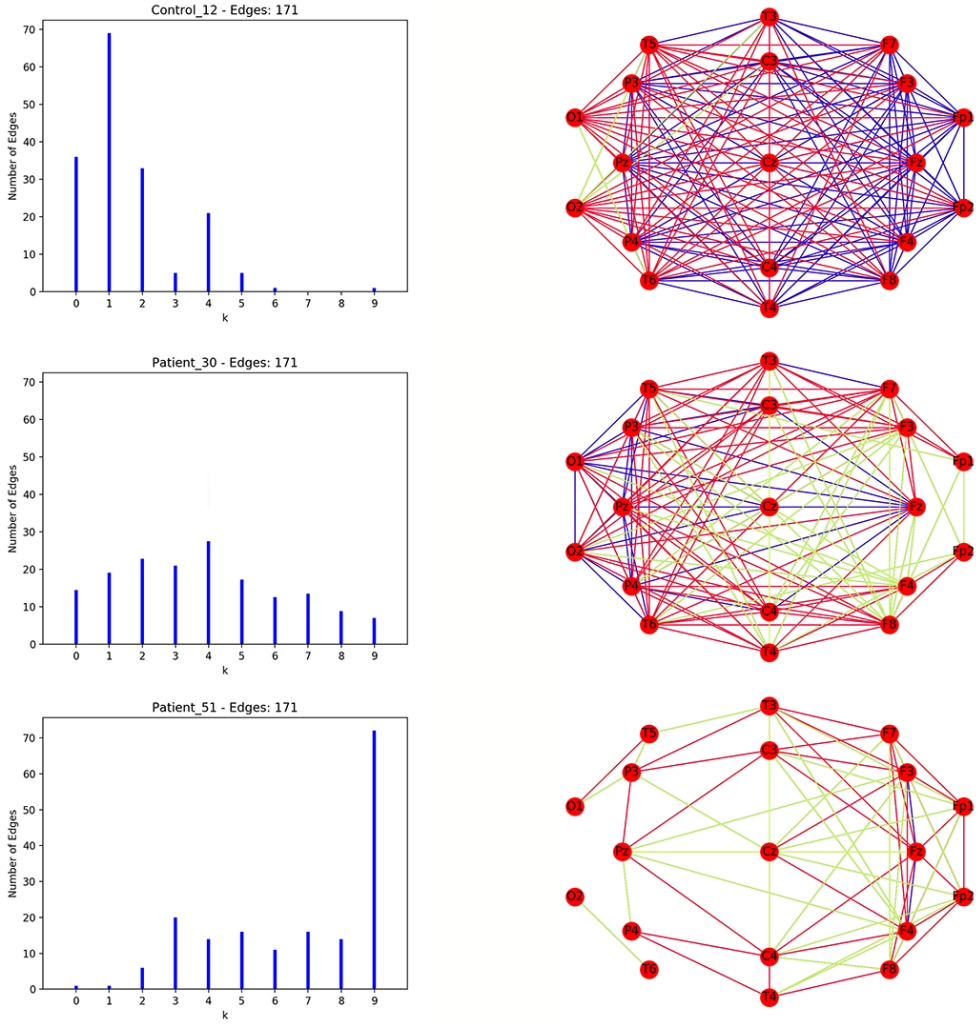


Fig. 3.1. Distributions of the edge weights and colored networks for the possible kinds of subjects into consideration. In particular, the first row is associated with a control subject, the second with a patient with MCI and the third with a patient with AD. In the distributions, k denotes the subrange number between min_E and max_E. In the networks, the disposal of nodes reflects the 10-20 system even if nodes are rotated 90 degrees clockwise. Observe that the control subject presents a high number of edges and most of them are blue; the corresponding distribution is biased towards left. The patient with MCI presents many edges and most of them are red; the corresponding distribution is balanced. The patient with AD presents a small number of edges and most of them are green; the corresponding distribution is biased towards right.

- CV denotes the set of the nodes of $\mathcal{CN}$. There is a node $v_i \in CV$ for each node of $\mathcal{N}_\pi$. A weight w_i is associated with v_i. It represents the number of cliques, which v_i is involved in. Formally speaking, let v_i be a node of $\mathcal{N}_\pi$ and let $\mathcal{C}_i$ be the set of the cliques of $\mathcal{N}_\pi$ which v_i is involved in (clearly $\mathcal{C}_i \subseteq \mathcal{C}$). Then CV is defined as:

$$CV = \{(v_i, w_i) | v_i \in V, w_i = |\mathcal{C}_i|\}$$

- CE represents the set of the edges of $\mathcal{CN}$. There is an edge $(v_i, v_j, w_{ij}) \in CE$ if the edge (v_i, v_j) is present in at least one clique of $\mathcal{C}$. w_{ij} denotes the number of cliques of $\mathcal{C}$, which (v_i, v_j) is involved in.

The edges of $\mathcal{CN}$ can be "colored" in an analogous way to the edges of $\mathcal{N}_\pi$. Also in this case, blue edges are the strongest ones, red edges have an intermediate strength and green edges are the weakest ones. Formally speaking:

$$CE = CE^b \cup CE^r \cup CE^g$$

Here:

- $CE^b = \{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE,\ w_{ij} > th_{rb}\}$;
- $CE^r = \{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE,\ (w_{ij} > th_{gr}) \wedge (w_{ij} \leq th_{rb})\}$;
- $CE^g = \{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE, w_{ij} \leq th_{gr}\}$.

Analogously to what we have seen for $\mathcal{N}_\pi$, we experimentally determined the values of th_{rb} and th_{gr}. In particular, we found that the best values for them are $th_{gr} = 4$ and $th_{rb} = 6$. We point out that clique network is very expressive from a visual point of view. Indeed, the color of an edge is an indicator of the strength of the connection between the corresponding brain areas, whereas the dimension of a node is an indicator of the connection degree of the corresponding brain area, and, ultimately, an indicator of its activity level.

In Figure 3.2, we report the clique networks corresponding to the EEGs of three patients at the time instants t_0 and t_1. Here, the dimension of a node is directly proportional to the associated weight. In this figure and in the following, we use the notation Patient X (MCI-MCI) - where X is a number - to denote a patient suffering from MCI at both t_0 and t_1. Analogously, Patient X (MCI-AD) indicates a patient with MCI at t_0 and AD at t_1. Finally, Patient X (AD-AD) represents a patient with AD at both t_0 and t_1.

Analogously to what we have done for colored networks, also in this case, in Table 3.2, we provide some quantitative measures characterizing the clique networks of Figure 3.2. Specifically, in this case, the considered measures are: *(i)* the total number of colored edges; *(ii)* the total number of blue (resp., red, green) edges; *(iii)* the percentage of colored edges against the total number of theoretically possible edges; *(iv)* the number of nodes with weights from 1 to 10. Even in this case, the quantitative values reported in this table fully confirm the qualitative analysis mentioned above.

In Appendix A.2, we report the pseudo-code for the construction of a clique network.

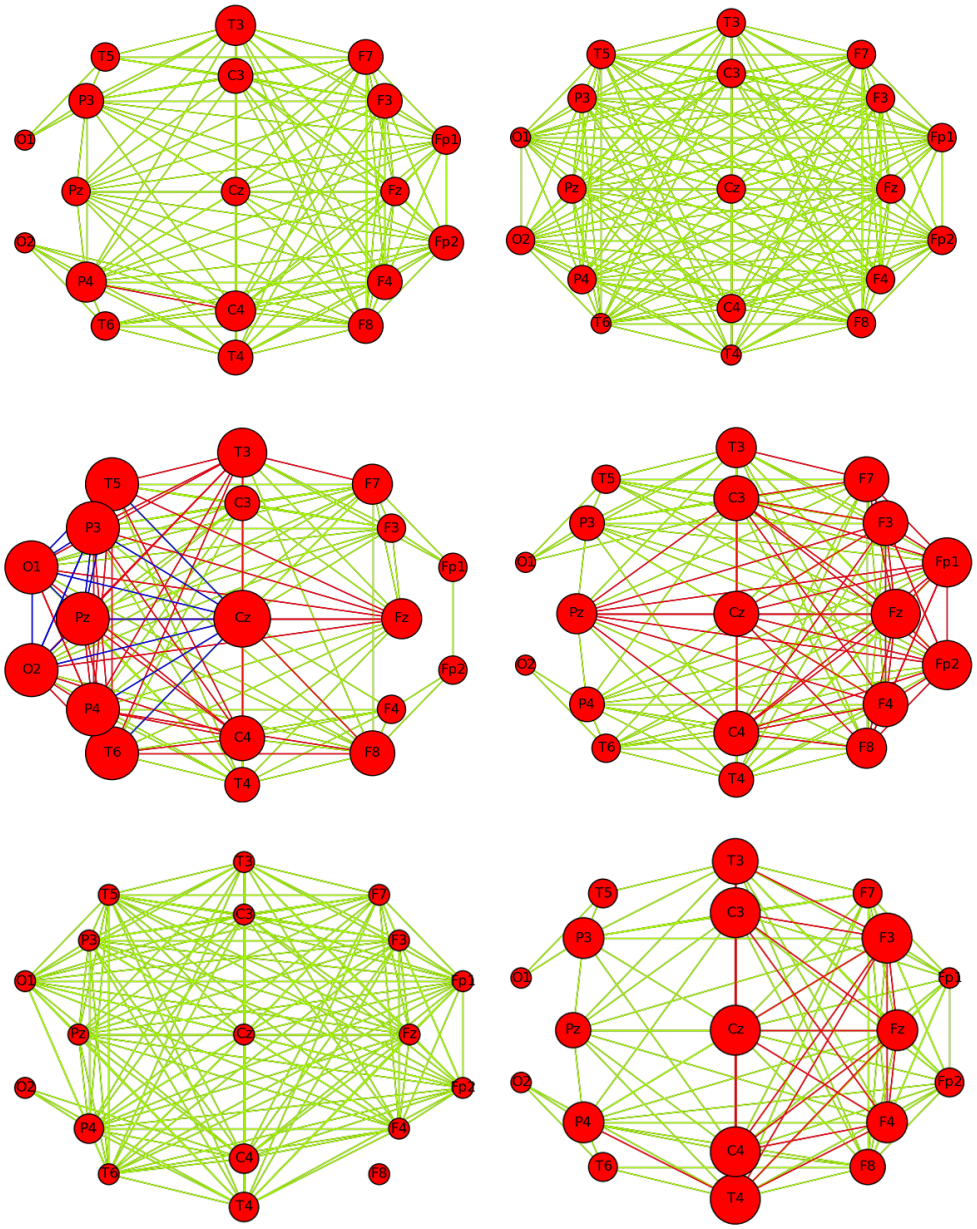


Fig. 3.2. The clique networks of Subjects 12 (Control Subject), 30 (MCI-MCI) and 51 (MCI-AD) at t_0 (on the left) and t_1 (on the right)

3.2.2 Connection Coefficient

As pointed out in the Introduction, one of the main features to investigate in neurodegenerative patients is the connection level of the brain areas. Previously, we introduced the concept of clique, which is one of the most powerful tools in network analysis for investigating the connection level of a network. Starting from cliques, it is possible to define a quantitative coefficient, which we call *connection coefficient*, capable of measuring the connectivity level of a network associated with an EEG.

This coefficient should take the following considerations into account:

Parameter	*Control 12 at t_0 (Control)*	*Control 12 at t_1 (Control)*	*Patient 30 at t_0 (MCI)*	*Patient 30 at t_1 (MCI)*	*Patient 51 at t_0 (MCI)*	*Patient 51 at t_1 (AD)*
Total number of colored edges	129	171	123	122	148	107
Total number of blue edges	0	0	23	0	0	0
Total number of red edges	1	0	40	48	0	21
Total number of green edges	128	171	60	74	148	86
Percentage of colored edges	75.4%	100%	71.9%	71.3%	86.5%	62.6%
Percentage of blue edges	0%	0%	13.6%	0%	0%	0%
Percentage of red edges	0.6%	0%	23.4%	28%	0%	12.3%
Percentage of green edges	74.8%	100%	35.1%	43.3%	86.5%	50.3%
Number of nodes whose weight is 0	0	0	0	0	1	0
Number of nodes whose weight is 1	2	19	0	2	15	3
Number of nodes whose weight is 2	6	0	4	2	3	4
Number of nodes whose weight is 3	8	0	1	3	0	3
Number of nodes whose weight is 4	3	0	3	3	0	4
Number of nodes whose weight is 5	0	0	2	6	0	5
Number of nodes whose weight is 6	0	0	1	3	0	0
Number of nodes whose weight is 7	0	0	6	0	0	0
Number of nodes whose weight is 8	0	0	2	0	0	0
Number of nodes whose weight is 9	0	0	0	0	0	0
Number of nodes whose weight is 10	0	0	0	0	0	0

Table 3.2. Quantitative results representing the networks of Figure 3.2

- Both the dimension and the number of cliques are important as connectivity indicators.
- The concept of clique is intrinsically exponential. In other words, a clique of dimension $n+1$ is exponentially more complex than a clique of dimension n.
- It is necessary to avoid the possible presence of outliers and noise. As a consequence, it is inappropriate to consider only the cliques with the maximum dimension. By contrast, it is more equilibrate to consider, in addition to them, the cliques with the sub-maximum and sub-sub-maximum dimension. On the other hand, it is unnecessary and time consuming to consider the other cliques because their contribution decreases exponentially against their dimension.

Starting from these considerations, we now define our connection coefficient. Let $\mathcal{N}_\pi = \langle V, E_\pi \rangle$ be the colored network associated with an EEG of $EEGSet$. Let $\mathcal{C}$ be the set of the cliques of $\mathcal{N}_\pi$ and let $dim(\cdot)$ be a function returning the dimension of a set of cliques, all of the same dimension, received in input. Then, it is possible to define: *(i)* the subset $\mathcal{C}_{M_1} \subseteq \mathcal{C}$ of the cliques with the maximum dimension; *(ii)* the subset $\mathcal{C}_{M_2} \subset \mathcal{C}$ of the cliques with the sub-maximum dimension; *(iii)* the subset $\mathcal{C}_{M_3} \subset \mathcal{C}$ of the cliques with the sub-sub-maximum dimension.

Finally, let $|\mathcal{C}_{M_1}|$, $|\mathcal{C}_{M_2}|$ and $|\mathcal{C}_{M_3}|$ be the cardinalities (i.e., the number of cliques) of $\mathcal{C}_{M_1}$, $\mathcal{C}_{M_2}$ and $\mathcal{C}_{M_3}$, respectively. Then, the connection coefficient $cc_{\mathcal{N}_\pi}$, associated with $\mathcal{N}_\pi$, is defined as:

$$cc_{\mathcal{N}_\pi} = \sum_{i=1}^{3} \left(\mid \mathcal{C}_{M_i} \mid \cdot 2^{dim(\mathcal{C}_{M_i})} \right)$$

This definition considers all the above observations in the most suitable way.

3.2.3 Sub-band Analysis

In the previous sections, we have always considered the complete EEG tracing. However, in the literature, it is well known that an EEG tracing can be separated in several sub-bands (e.g., α, β, δ and θ) whose analysis can provide significant information in several neurological disorders. For instance, in the past, it was shown that the sub-bands δ and θ can help in investigating the conversion from MCI to AD [118, 64]. For this reason, we decided to extend all the previous analysis from the overall tracing to the ones of the sub-bands α, β, δ and θ. In this section, we illustrate this extension and the most important results we have obtained from it.

Preliminarily, we must introduce further support data structures and parameters. Specifically, let eeg be a generic EEG of $EEGSet$. Starting from eeg, it is possible to define four further tracings, namely eeg^{α}, eeg^{β}, eeg^{δ} and eeg^{θ}, referred to the sub-bands α, β, δ and θ.

In Section 3.2.1, we have defined the network $\mathcal{N} = \langle V, E \rangle$ corresponding to eeg. In an analogous way, it is possible to define the networks:

$$\mathcal{N}^{\alpha} = \langle V, E^{\alpha} \rangle \qquad \mathcal{N}^{\beta} = \langle V, E^{\beta} \rangle \qquad \mathcal{N}^{\delta} = \langle V, E^{\delta} \rangle \qquad \mathcal{N}^{\theta} = \langle V, E^{\theta} \rangle$$

Here, V is the set of nodes, which coincides with the nodes of $\mathcal{N}$. E^{α} (resp., E^{β}, E^{δ}, E^{θ}) represents the set of the edges of $\mathcal{N}^{\alpha}$ (resp., $\mathcal{N}^{\beta}$, $\mathcal{N}^{\delta}$, $\mathcal{N}^{\theta}$). Each edge of E^{α} (resp., E^{β}, E^{δ}, E^{θ}), connecting the nodes v_i and v_j, has the form (v_i, v_j, w_{ij}), where w_{ij} is a measure of the "distance" between v_i and v_j in $\mathcal{N}^{\alpha}$ (resp., $\mathcal{N}^{\beta}$, $\mathcal{N}^{\delta}$, $\mathcal{N}^{\theta}$). As seen in Section 3.2.1, this "distance" is an indicator of the disconnection level of v_i and v_j, and each measure representing this feature could be adopted in our model. Analogously to the overall tracing, in the experiments associated with this research, we adopted the Permutation Disalignment Index [292]. As a consequence, for the edge $(v_i, v_j, w_{ij}) \in E^{\alpha}$ (resp., E^{β}, E^{δ}, E^{θ}), w_{ij} is equal to the average value of PDI in eeg^{α} (resp., eeg^{β}, eeg^{δ}, eeg^{θ}).

Beside $\mathcal{N}^{\alpha}$, $\mathcal{N}^{\beta}$, $\mathcal{N}^{\delta}$ and $\mathcal{N}^{\theta}$, it is possible to define:

- the colored networks $\mathcal{N}_{\pi}^{\alpha} = \langle V, E_{\pi}^{\alpha} \rangle$ (resp., $\mathcal{N}_{\pi}^{\beta}$, $\mathcal{N}_{\pi}^{\delta}$, $\mathcal{N}_{\pi}^{\theta}$), corresponding to eeg^{α} (resp., eeg^{β}, eeg^{δ}, eeg^{θ}), by extending to this tracing what we have already done in Section 3.2.1 for the overall tracing;
- the connection coefficient $cc_{\mathcal{N}_{\pi}^{\alpha}}$ (resp., $cc_{\mathcal{N}_{\pi}^{\beta}}$, $cc_{\mathcal{N}_{\pi}^{\delta}}$, $cc_{\mathcal{N}_{\pi}^{\theta}}$), corresponding to eeg^{α} (resp., eeg^{β}, eeg^{δ}, eeg^{θ}), by extending to this tracing what we have already done in Section 3.2.2 for the overall tracing.

3.2.4 Conversion Coefficient

We have introduced the connection coefficient and we have shown that it is well suited for determining the connection degree of a network and, in our case, of the brain. In

this task, this parameter presents a better performance than clustering coefficient that is the parameter adopted in classical Network Analysis for this purpose. It also proved to be adequate to verify the conversion from MCI to AD. Finally, its adoption in sub-bands δ and θ proved to be well suited to predict the same conversion.

All these results, in the whole, suggest that, in order to quantitatively predict the conversion from MCI to AD, it is reasonable to define a new coefficient (which we call *conversion coefficient*) capable of detecting the conversion of a patient from MCI to AD more exactly, by taking the connection coefficient relative to all these three tracings into account.

The conversion coefficient can be defined as follows: let eeg be an EEG of $EEGSet$, let $\mathcal{N}_\pi$ (resp., $\mathcal{N}_\pi^\delta$, $\mathcal{N}_\pi^\theta$) be the corresponding colored network associated with the overall tracing (resp., the sub-bands δ and θ) of eeg, let $cc^0_{\mathcal{N}_\pi}$, $cc^0_{\mathcal{N}_\pi^\delta}$, $cc^0_{\mathcal{N}_\pi^\theta}$ (resp., $cc^1_{\mathcal{N}_\pi}$, $cc^1_{\mathcal{N}_\pi^\delta}$, $cc^1_{\mathcal{N}_\pi^\theta}$) be the corresponding connection coefficients at t_0 (resp., t_1). The conversion coefficient $conv_{eeg}$, corresponding to eeg, is defined as:

$$conv_{eeg} = \frac{1}{3} \cdot \left(\frac{cc^1_{\mathcal{N}_\pi} - cc^0_{\mathcal{N}_\pi}}{cc^0_{\mathcal{N}_\pi}} + \frac{cc^1_{\mathcal{N}_\pi^\delta} - cc^0_{\mathcal{N}_\pi^\delta}}{cc^0_{\mathcal{N}_\pi^\delta}} + \frac{cc^1_{\mathcal{N}_\pi^\theta} - cc^0_{\mathcal{N}_\pi^\theta}}{cc^0_{\mathcal{N}_\pi^\theta}} \right)$$

In other words, the conversion coefficient $conv_{eeg}$ of an electroencephalogram eeg considers the variations of the connection coefficients $cc_{\mathcal{N}_\pi}$, $cc_{\mathcal{N}_\pi^\delta}$ and $cc_{\mathcal{N}_\pi^\theta}$ associated with the overall tracing and with the tracings corresponding to the sub-bands δ and θ. All these contributions are taken with the same weight.

3.2.5 Network Motifs

In this section, we aim at investigating the possible presence of motifs characterizing patients with MCI from patients with AD, and vice versa.

As a matter of fact, motifs have been already investigated and used in past approaches adopting network analysis (see, for instance, [312, 414, 339]). In this scenario, they are considered as [312]:

> "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks".

In our approach, we use motifs in a completely different way. Indeed, we do not examine a unique complex network to find patterns frequently repeated therein. By contrast, we search for patterns appearing frequently in the networks corresponding to the tracing segments of patients with MCI (resp., AD) and being absent in the ones of patients with AD (resp., MCI), thus characterizing the patients with MCI (resp., AD) from the ones with AD (resp., MCI).

First, we must formalize our concept of motif. Specifically, let $EEGSet$ be a set of EEGs, let $MCISet$ (resp., $ADSet$) be the subset of $EEGSet$ corresponding to

patients with MCI (resp., AD). Let $NSet_M$ (resp., $NSet_A$) be the set of colored networks corresponding to the EEGs of $MCISet$ (resp., $ADSet$). Let $\mathcal{C}_M$ (resp., $\mathcal{C}_A$) be the set of the cliques of $NSet_M$ (resp., $NSet_A$) and let $\mathcal{T}_M$ (resp., $\mathcal{T}_A$) be the set of totally connected triads of $\mathcal{C}_M$ (resp., $\mathcal{C}_A$)[4]. Finally, let t be a generic triad. We call $nocc_M$ (resp., $nocc_A$) the number of occurrences of t in $NSet_M$ (resp., $NSet_A$).

After having defined all support data structures and parameters, we are able to describe our motif extraction approach. It consists of two main steps, the former devoted to the extraction of basic motifs and the latter conceived for the construction of derived ones. In this section, we illustrate the extraction of basic motifs. Preliminarily, it is necessary to specify what a basic motif is in our context. Specifically:

> Let t be a totally connected triad of $NSet_M$. If: (1) t is *absent* in the networks of $NSet_A$ and is *frequent* in the networks of $NSet_M$, or (2) t is *very rare* in the networks of $NSet_A$ and *very frequent* in the networks of $NSet_M$, then t is a motif characterizing patients with MCI from patients with AD.

To really extract basic motifs, it is necessary to specify a quantitative definition of this rule. To carry out this task, it is preliminarily necessary to associate numeric values with the concepts of *very rare*, *frequent* and *very frequent*. For this purpose, we can define the following thresholds:

$$th_{VR} = \alpha_{VR} \cdot |NSet_A| \qquad th_F = \alpha_F \cdot |NSet_M|$$
$$th_{VF} = \alpha_{VF} \cdot |NSet_M|$$

We experimentally set the values of α_{VR}, α_F and α_{VF} to 0.10, 0.25 and 0.40, respectively. We chose these values as the default ones of our approach. In fact, they proved to be the most "equilibrate" (i.e., neither extremely permissive nor extremely restrictive) ones.

Therefore, let $t \in \mathcal{T}_M$ be a totally connected triad of $NSet_M$ and let $nocc_M$ (resp., $nocc_A$) be the number of occurrences of t in $NSet_M$ (resp., $NSet_A$). If:

- (1) $(nocc_A = 0) \wedge (nocc_M \geq th_F)$, or
- (2) $(nocc_A \leq th_{VR}) \wedge (nocc_M \geq th_{VF})$

then t is a basic motif characterizing patients with MCI from patients with AD.

In a dual fashion, it is possible to define the basic motifs characterizing patients with AD from patients with MCI. Also in this case, we experimentally set the values of α_{VR}, α_F and α_{VF} to 0.10, 0.25 and 0.40, respectively.

[4] We recall that a triad is a subnetwork consisting of three nodes. The totally connected triad is considered the most stable structure in network analysis. A totally connected triad can be considered as a clique of dimension 3.

In the following, we indicate by $\mathcal{M}_M$ (resp., $\mathcal{M}_A$) the set of motifs extracted starting from the triads of $NSet_M$ (resp., $NSet_A$).

Observe that a motif is not only an indicator of the tracing segments of the EEGs of patients with MCI (or with AD). As a matter of fact, it is much more. Indeed, it allows us to characterize the behavior of the brain areas of patients with MCI (resp., AD) from patients with AD (resp., MCI). For instance, it denotes what brain areas are most connected (and, therefore, most active) in patients with MCI before converting to AD (resp., in patients that converted from MCI to AD).

Once basic motifs have been extracted, and a first version of $\mathcal{M}_M$ and $\mathcal{M}_A$ has been obtained, it is possible to construct derived (and, possibly, much more complex and significant) motifs starting from them.

Our approach constructs new derived motifs starting from the already known ones. For this purpose, it uses nodes common to two or more known motifs as "junction points". Formally speaking, let $m_i = \langle V_i, E_i \rangle$ and $m_j = \langle V_j, E_j \rangle$ be two motifs of $\mathcal{M}_M$ such that $V_i \cap V_j \neq \emptyset$. Then, it is possible to construct a candidate motif by computing the union of the nodes and the edges of m_i and m_j:

$$m_{ij} = \langle V_i \cup V_j, E_i \cup E_j \rangle$$

Once m_{ij} has been constructed, analogously to what we have seen for basic motifs, it is necessary to evaluate $nocc_M$ and $nocc_A$[5]. If, for these parameters, condition (1) or condition (2) for the extraction of basic motifs hold, then m_{ij} can be added to $\mathcal{M}_M$, i.e., $\mathcal{M}_M = \mathcal{M}_M \cup \{m_{ij}\}$.

The addition of a new motif in $\mathcal{M}_M$ could make possible the construction of new candidate motifs. As a consequence, the enrichment process of $\mathcal{M}_M$ is iterative and terminates when, during an iteration, no new motif is added to $\mathcal{M}_M$. In an analogous fashion, the derived motifs of $\mathcal{M}_A$ can be extracted. In Appendix A.3, we report the pseudo-code for the computation of motifs.

3.3 Results

3.3.1 Testbed

We enrolled seven patients with AD and eight patients with MCI monitored at the IRCCS Centro Neurolesi Bonino Pulejo of Messina (Italy), within a three-month follow-up program. The main characteristics of these patients are reported in Table 3.3.

[5] Clearly, for derived motifs, $nocc_M$ and $nocc_A$ refer to the number of occurrences of motifs, instead of triads.

Patient ID	*Age*	*Gender*	*Diagnosis at* t_0	*Diagnosis at* t_1
pt_03	68	M	MCI	AD
pt_23	84	F	MCI	MCI
pt_30	69	M	MCI	MCI
pt_41	78	M	MCI	MCI
pt_51	71	F	MCI	AD
pt_57	83	M	MCI	MCI
pt_71	79	F	MCI	AD
pt_72	65	F	MCI	MCI
pt_31	74	M	AD	AD
pt_54	83	F	AD	AD
pt_64	74	F	AD	AD
pt_65	76	M	AD	AD
pt_76	79	F	AD	AD
pt_86	83	F	AD	AD
pt_87	78	F	AD	AD

Table 3.3. Main characteristics of the patients enrolled for our experiments

Every subject signed an informed consent form, in agreement with a clinical protocol approved by the Ethical Committee. We also enrolled eighteen control subjects. The diagnostic procedure followed the guidelines of the Diagnostic and Statistical Manual of Mental Disorders (fifth edition, DSM-5) [37] and consisted of a full cognitive and clinical assessment, carried out by a multidisciplinary team of neurologists, psychologists, psychiatrists and EEG experts. Each patient was evaluated at baseline (time t_0) and then again three months later (time t_1). The patients were evaluated neuroradiologically, in order to rule out other clinical conditions, like brain lesions, which might have caused cognitive impairment. Current medical treatment (particularly cholinesterase inhibitors - ChEis, Memantine, anti-depressants, anti-psychotics and anti-epileptic drugs) was also taken into account in AD patients. MCI subjects were not under medical treatment. Furthermore, we also had 18 EEGs of control subjects.

The EEGs were recorded according to the 10-20 International System (19 channels), with 1024 Hz sampling rate. A 50 Hz notch filter was used, with linked earlobe (A1-A2) reference. The EEG recordings were performed in a comfortable resting state. The patients kept their eyes closed but remained awake. The EEG was band-pass filtered at 0.5-32 Hz through the Matlab toolbox *EEGLab* (`https://sccn.ucsd.edu/eeglab/`) [126]. EEG preprocessing was fully carried out in Matlab (The MathWorks, Inc., Natick, MA, USA). After filtering, the artifactual segments in the EEG recordings were manually detected by the EEG experts and the artifactual epochs were discarded. The average time length of the recordings, after artifact cancellation, is 5.44 *mins*. After that, the four major EEG rhythms, i.e., α, β, δ and θ were extracted from the EEG signals. In this way, a n-channels EEG recording was eventually split into 4 n-channels sub-band EEG recordings: EEG^{α}, EEG^{β}, EEG^{δ}, EEG^{θ}. Each sub-band of the EEG was then downsampled to 256 Hz. Every record-

ing of the sub-bands was partitioned into 5 s non-overlapping windows, and analyzed window by window.

On the basis of the diagnosis at times t_0 and t_1, the patients into examination were partitioned in three groups, namely: *(i)* patients with MCI at t_0 that were still diagnosed MCI at t_1; *(ii)* patients with AD at t_0 that remained with AD at t_1; *(iii)* patients with MCI at t_0 that converted to AD at t_1.

As pointed out in the Introduction, we have striven to (at least partially) face the issue of the narrowness of the set of available patients. For this purpose, we realized a simulator aimed to construct virtual control subjects and virtual patients with MCI or AD. The simulator behaves as follows:

- It receives a set $ASet^{CS}$ (resp., $ASet^{MCI}$, $ASet^{AD}$) of matrices. Each element of this set represents the adjacency matrix of the complex network associated with the EEG of a control subject (resp., a patient with MCI, a patient with AD). The set of real control subjects (resp., patients with MCI, patients with AD) from which we constructed $ASet^{CS}$ (resp., $ASet^{MCI}$, $ASet^{AD}$) consisted of the 50% of the control subjects (resp., patients with MCI, patients with AD) at our disposal, selected at random. In fact, as we will see below, the other 50% of control subjects (resp., patients with MCI, patients with AD) were necessary for testing our simulator.
- It constructs a new adjacency matrix $\overline{A^{CS}}$ (resp., $\overline{A^{MCI}}$, $\overline{A^{AD}}$) whose generic element $\overline{A^{CS}}[i,j]$ (resp., $\overline{A^{MCI}}[i,j]$, $\overline{A^{AD}}[i,j]$) represents the mean of the (i,j) elements of the matrices of $ASet^{CS}$ (resp., $ASet^{MCI}$, $ASet^{AD}$).
- It computes the standard deviation σ^{CS} (resp., σ^{MCI}, σ^{AD}) of the elements of $\overline{A^{CS}}$ (resp., $\overline{A^{MCI}}$, $\overline{A^{AD}}$).
- It constructs the set $\widehat{ASet^{CS}}$ (resp., $\widehat{ASet^{MCI}}$, $\widehat{ASet^{AD}}$) of the adjacency matrices representing the complex networks associated with the EEGs of virtual control subjects (resp., patients with MCI, patients with AD). In particular, the generic element $\widehat{A}[i,j]$ of a matrix of $\widehat{ASet^{CS}}$ (resp., $\widehat{ASet^{MCI}}$, $\widehat{ASet^{AD}}$) is obtained by perturbing the corresponding element $\overline{A^{CS}}$ (resp., $\overline{A^{MCI}}$, $\overline{A^{AD}}$) of a random value comprising between $-\frac{1}{2}\sigma^{CS}$ (resp., $-\frac{1}{2}\sigma^{MCI}$, $-\frac{1}{2}\sigma^{AD}$) and $\frac{1}{2}\sigma^{CS}$ (resp., $\frac{1}{2}\sigma^{MCI}$, $\frac{1}{2}\sigma^{AD}$).

After having obtained the three sets $\widehat{ASet^{CS}}$, $\widehat{ASet^{MCI}}$, $\widehat{ASet^{AD}}$, it was necessary to couple the corresponding matrices appropriately in such a way as to represent virtual control subjects (having an element of $\widehat{ASet^{CS}}$ at t_0 and another one of the same set at t_1), virtual patients with MCI at both t_0 and t_1 (therefore, having an element of $\widehat{ASet^{MCI}}$ at t_0 and another one of the same set at t_1), and patients with MCI at t_0 and with AD at t_1, and, finally, patients with AD at both t_0 and t_1. Each of

the four sets constructed above consisted of 27 elements. After this, by following the holdout technique, for each of the four groups mentioned above, we chose 18 elements to train our approach and 9 elements to test it. After having verified the adequacy of our approach on virtual people, we tested it on the 50% of the real people not used for constructing the virtual models, in such a way as to verify its suitability on real patients. We applied this technique first to evaluate the connection coefficient on the overall EEG tracing, then to test the same coefficient on the four EEG sub-bands and, finally, to evaluate the conversion coefficient.

Before discussing the "adequacy" of our approach, a discussion about the enrollment of patients in neurological tests is in order. Nowadays it is still very difficult to keep MCI and AD subjects and their caregivers actively involved in the follow-up programs. On the other side, these programs are strictly necessary to develop biomarkers for the objective quantification of the degeneration degree of cortical electrical connectivity caused by dementia. Many subjects do not fulfil the timing of the periodic assessments. This is often due to the difficulties caused by the disease itself. This means that many recruited subjects must be later excluded from the analysis because their EEGs were not recorded following the predetermined scheduling, which implies that their inclusion would not allow the construction of a dataset with homogeneous characteristics. As a result, there are only a few longitudinal studies in which the EEG of the subjects has been recorded and evaluated twice over time. To the best of our knowledge, the largest sample ever analyzed (143 MCI subjects) was constructed within a multicentric study described in [85]. In this paper, the authors introduced a methodology, named Implicit Function As Squashing Time (IFAST), based on artificial neural networks. IFAST succeeded to predict the conversion from amnestic MCI to AD with a 85.98% accuracy in a 1-year follow-up study. Later, this methodology was improved; however, it has been so far tested only on a classification study concerning cross-sectional MCI vs AD.

Some other follow-up studies were carried out, but the EEG was recorded and assessed only at baseline (i.e., at t_0) and was later interpreted on the basis of the new diagnosis formulated at time t_1. In particular, [195] examined 35 amnestic MCI subjects whose EEGs were recorded at time t_0. Then, they retrospectively classified these EEGs according to the diagnosis reformulated at time t_1. The features were extracted through a Phase Lag Index (PLI)-based connectivity analysis. [323] analyzed the correlation between higher alpha3/alpha2 frequency power, cortical decay and perfusion rate with conversion to AD in a group of 76 subjects diagnosed as MCI patients at time t_0 and, then, re-evaluated at time t_1. [176] recruited 205 nondemented amyloid positive subjects (142 of them were MCI), and computed peak frequencies and relative power in the four major sub-bands (δ, θ, α, β). Then, they retrospectively

evaluated the relationship between normalized EEG measures and the probability of conversion to AD. The study proposed by [368] included 86 MCI subjects. These authors introduced a Neurophysiological Biomarker Toolbox, based on β band features, to predict the conversion from MCI to AD.

All the aforementioned studies consisted in a retrospective cross-sectional classification between groups of subjects. They do not perform the longitudinal quantification of changes in the EEGs of the same subject, which is the only way to find possible correlations between changes in the characteristics of EEG signals and/or physiological changes caused by the progression of the disease.

After this premise, we can proceed to quantitatively measure the "adequacy" of our approach, we adopted the parameters generally used in the literature for this purpose (see [191] for all details). In particular, let *pos* be the number of positives in a clinical analysis (in our case, the number of patients converting from MCI to AD in real life), let *t_pos* be the number of true positives (in our case, the number of patients converting from MCI to AD in real life and correctly detected by the connection coefficient), let *f_pos* be the number of false positives, let *neg* be the number of negatives and, finally, let *t_neg* be the number of true negatives. Starting from these parameters, it is possible to define:

- *sensitivity*, or *true positive rate*, as the proportion of positives correctly identified by the approach to evaluate: $sensitivity = \frac{t\_pos}{pos}$;
- *specificity*, or *true negative rate*, as the proportion of negatives correctly identified by the approach to evaluate: $specificity = \frac{t\_neg}{neg}$;
- *precision*, as the proportion of subjects labeled as positives by the approach to evaluate and being really positives: $precision = \frac{t\_pos}{t\_pos + f\_pos}$.

Clearly, in this medical context, sensitivity is much more important than specificity and precision.

As a final remark, we performed a comparative evaluation of our connection and conversion coefficients against clustering coefficient, which is much simpler and is the classical parameter adopted in network analysis to evaluate the connection level of a network.

3.3.2 Training of the proposed approach

First, we decided to perform a preliminary, yet rough, verification of the capability of our EEG generator to produce plausible results. For this purpose, we computed the average minimum weight, the average maximum weight and the average mean weight for the following sets: *(i)* 50% of real EEGs (control subjects, patients with MCI, patients with AD) used to "train" the EEG generator; *(ii)* virtual EEGs produced

through our generator and used to train our approach; *(iii)* virtual EEGs produced through our generator and used to test our approach; *(iv)* 50% of real EEGs used to test our approach. Obtained results are reported in Table 3.4.

Set of persons	*Avg Min Weight*	*Avg Mean Weight*	*Avg Max Weight*
Real control subjects for generator training	1.2852	1.8534	3.0923
Real control subjects for approach testing	1.2114	1.8355	3.0954
Virtual control subjects for approach training	1.1887	1.8543	2.9367
Virtual control subjects for approach testing	1.1511	1.8446	2.8912
Real patients with MCI for generator training	1.3612	2.0812	3.0224
Real patients with MCI for approach testing	1.2729	1.8854	2.7689
Virtual patients with MCI for approach training	1.2723	1.8838	2.4678
Virtual patients with MCI for approach testing	1.2863	1.8856	2.4643
Real patients with AD for generator training	1.2867	2.0243	2.9498
Real patients with AD for approach testing	1.3412	2.0976	3.0657
Virtual patients with AD for approach training	1.2643	2.0385	2.9564
Virtual patients with AD for approach testing	1.2712	2.0501	2.9504

Table 3.4. Average minimum weight, average mean weight and average maximum weight for the sets of interest

From the analysis of this table we can observe that they appear plausible, similar to the corresponding real ones and, at the same time, present a reasonable heterogeneity. For instance, the maximum variation of the average minimum (resp., mean, maximum) weight is 7.90% (resp., 9.62%, 18.24%).

After this verification, we trained our approach for making it able to detect the conversion from MCI to AD. With regard to this task, we found that a decrease of the connection coefficient higher than 80% is a potentially good indicator of the conversion phenomenon. We found the identical threshold value also for the conversion coefficient.

3.3.3 Testing of the proposed approach

The first test that we performed regarded the connection coefficient's capability of detecting the conversion of a patient from MCI to AD.

First we operated on virtual EEGs. As previously specified, we considered 27 virtual patients with MCI at both t_0 and t_1, 27 virtual patients with AD at both t_0 and t_1 and 27 virtual patients with MCI at t_0 that converted to AD at t_1. Obtained results are shown in the first row of Table 3.5. Then, we considered real people and operated exactly as in the previous test. Obtained results are reported in the second row of Table 3.5. The analysis of this table shows that the connection coefficient appears a good parameter for predicting the conversion from MCI to AD. Sensitivity, specificity and precision obtained by this coefficient are very high, even if improvable, both for virtual patients and for real ones. Interestingly, the values obtained for real patients are higher than the ones returned for virtual patients.

Set	Sensitivity	Specificity	Precision
Virtual patients	0.94	0.91	0.72
Real patients	1.00	0.91	0.75

Table 3.5. Sensitivity, specificity and precision of the connection coefficient associated with overall EEGs

The second test was analogous to the first one, but it regarded the sub-bands of EEGs, instead of the overall tracing. The corresponding results are reported in Tables 3.6 and 3.7. These tables show that δ and θ sub-bands are very adequate for investigating the conversion of a patient from MCI to AD. This result is in line with the ones obtained by [118, 64]. Also for these sub-bands, real patients behave better than virtual ones. α and β sub-bands, instead, do not present particularly satisfying results. For all these reasons, we decided to not consider these two sub-bands in the computation of the conversion coefficient.

Set	Sensitivity	Specificity	Precision
Virtual patients (α sub-band)	0.75	0.94	0.71
Virtual patients (β sub-band)	0.85	0.80	0.72
Virtual patients (δ sub-band)	0.94	0.95	0.69
Virtual patients (θ sub-band)	0.92	0.97	0.54

Table 3.6. Sensitivity, specificity and precision of the connection coefficient associated with the sub-bands of EEGs (virtual patients)

Set	Sensitivity	Specificity	Precision
Real patients (α sub-band)	0.67	0.91	0.67
Real patients (β sub-band)	0.80	0.80	0.67
Real patients (δ sub-band)	1.00	1.00	0.75
Real patients (θ sub-band)	1.00	1.00	0.60

Table 3.7. Sensitivity, specificity and precision of the connection coefficient associated with the sub-bands of EEGs (real patients)

The next test regarded the conversion coefficient's capability of detecting the conversion of a patient from MCI to AD. For this purpose, we operated in an analogous way to what we have seen for the connection coefficient, i.e., first we considered virtual EEGs and, then, real ones. Obtained results are reported in Table 3.8.

Set	Sensitivity	Specificity	Precision
Virtual patients	0.95	0.94	0.92
Real patients	1.00	1.00	1.00

Table 3.8. Sensitivity, specificity and precision of the conversion coefficient

As shown in this table, the values of sensitivity, specificity and precision returned by conversion coefficient are extremely high for virtual patients and maximum for real ones. Again, real patients behave better than virtual ones.

As for the comparison between the connection and the clustering coefficients in distinguishing control subjects from patients with MCI and patients with AD, from the analysis of Table 3.9 we observe that:

- The average connection coefficient of virtual (resp., real) patients with MCI decreases of 14.45% (resp., 11.32%) w.r.t. the corresponding value of virtual (resp., real) control subjects. Instead, the average clustering coefficient of virtual (resp., real) patients with MCI decreases of 2.46% (resp., 2.39%) w.r.t. the corresponding value of virtual (resp., real) control subjects.
- The average connection coefficient of virtual (resp., real) patients with AD decreases of 75.77% (resp., 69.63%) w.r.t. the corresponding value of virtual (resp., real) patients with MCI. Instead, the average clustering coefficient of virtual (resp., real) patients with AD decreases of 15.16% (resp., 12.81%) w.r.t. the corresponding value of virtual (resp., real) patients with MCI.

These values clearly evidence that the connection coefficient is much better than the clustering coefficient in distinguishing control subjects, patients with MCI and patients with AD. As a consequence, even if the computation of this coefficient is more expensive than the one of the clustering coefficient, this is balanced by its much better capability of distinguishing the states of a person.

3.3.4 Comparison between Connection and Clustering coefficients

As previously pointed out, in Social Network Analysis, the most commonly used parameter for evaluating the connection level of a network is clustering coefficient. This coefficient is simpler to compute than the connection and the conversion coefficients. As a consequence, the adoption of these last ones makes sense only if they provide more accurate results. To verify if this happens, we performed some tests.

The first one aimed at computing the average connection coefficient and the average clustering coefficient for virtual and real control subjects, patients with MCI and patients with AD. The obtained results are reported in Table 3.9.

The second test aimed at comparing the capability of the conversion and the clustering coefficients in determining the conversion of a patient from MCI to AD. In Table 3.8, we report sensitivity, specificity and precision of the conversion coefficient in carrying out this task.

We performed the same analysis for clustering coefficient. In this case, we experimentally set to 80% the percentage of the decrease of the clustering coefficient

Set	Average Connection Coefficient	Average Clustering Coefficient
Virtual control subjects	232523	0.9675
Virtual patients with MCI	198785	0.9422
Virtual patients with AD	48223	0.7889
Real control subjects	226169	0.9592
Real patients with MCI	200548	0.9363
Real patients with AD	60904	0.8164

Table 3.9. Average connection coefficient and average clustering coefficient for all the sets of virtual and real people of interest

necessary for saying that a patient converted from MCI to AD. The corresponding sensitivity, specificity and precision are reported in Table 3.10.

Set	Sensitivity	Specificity	Precision
Virtual patients	0.77	0.71	0.68
Real patients	0.82	0.84	0.75

Table 3.10. Sensitivity, specificity and precision of the clustering coefficient

The analysis of Tables 3.8 and 3.10 allows us to point out that conversion coefficient returned much better results than clustering coefficient. In fact, for virtual (resp., real) patients, sensitivity, specificity and precision increase of 26.31%, 32.86% and 34.78% (resp., 21.95%, 19.05% and 33.33%) if the conversion coefficient is adopted in place of the clustering coefficient.

These two tests allow us to conclude that, even if our coefficients are more complex than the clustering coefficient, they can provide much better results and, therefore, are worthwhile to be adopted.

3.3.5 Network Motifs

The basic motifs belonging to $\mathcal{M}_M$ derived by our approach are reported in Table 3.11.

On the top of Figure 3.3, we represent two basic motifs belonging to $\mathcal{M}_M$, obtained by applying our approach to the EEGs of the patients at our disposal.

With the current values of α_{VR}, α_F and α_{VF}, we did not extract any motif belonging to $\mathcal{M}_A$. This is in line with the results shown in Sections 3.3.4, where we have seen that the networks corresponding to patients with AD are much less connected than the ones corresponding to patients with MCI. However, if the human expert wants to be more "permissive", she/he can decrease the values of α_F and α_{VF} and can increase the value of α_{VR} w.r.t. the default ones specified above. In this case, she/he could find basic motifs also in $\mathcal{M}_A$.

On the bottom of Figure 3.3, we show the most significant derived motifs extracted by our approach. In order to provide a quantitative evaluation of derived motifs (which implies characterizing the tracing segments of patients with MCI from patients

Condition (1)	Condition (2)
$[Fp1, Fp2, O2]$; $[Fp1, F3, O2]$; $[Fp1, Fz, O2]$	$[Fz, C3, O2]$
$[Fp1, F4, O2]$; $[Fp1, F8, O2]$; $[Fp1, C3, O2]$	$[Fz, Cz, O2]$
$[Fp1, Cz, O2]$; $[Fp1, C4, O2]$; $[Fp1, T4, O2]$	$[Fz, C4, O2]$
$[Fp1, Pz, O2]$; $[Fp1, P4, O2]$; $[Fp1, T6, O2]$	$[Fz, T4, O2]$
$[Fp2, F3, O2]$; $[Fp2, Fz, O2]$; $[Fp2, F4, O2]$	$[Fz, Pz, O2]$
$[Fp2, F8, O2]$; $[Fp2, C3, O2]$; $[Fp2, Cz, O2]$	$[Fz, P4, O2]$
$[Fp2, C4, O2]$; $[Fp2, T4, O2]$; $[Fp2, Pz, O2]$	$[Fz, T6, O2]$
$[Fp2, P4, O2]$; $[Fp2, T6, O2]$; $[F7, F3, O2]$	$[C3, Cz, O2]$
$[F7, Fz, O2]$; $[F7, Cz, O2]$; $[F7, C4, O2]$	$[C3, C4, O2]$
$[F7, P4, O2]$; $[F3, Fz, O2]$; $[F3, F4, O2]$	$[C3, Pz, O2]$
$[F3, F8, O2]$; $[F3, T3, O2]$; $[F3, C3, O2]$	$[C3, P4, O2]$
$[F3, Cz, O2]$; $[F3, C4, O2]$; $[F3, T4, O2]$	$[C3, T6, O2]$
$[F3, T5, O2]$; $[F3, P3, O2]$; $[F3, Pz, O2]$	
$[F3, P4, O2]$; $[F3, T6, O2]$; $[F3, O1, O2]$	
$[Fz, F4, O2]$; $[Fz, F8, O2]$; $[Fz, T3, O2]$	
$[F4, C3, O2]$; $[F8, C3, O2]$; $[F8, P3, O2]$	
$[T3, C4, O2]$	

Table 3.11. The basic motifs belonging to $\mathcal{M}_M$ derived by applying condition (1) and condition (2)

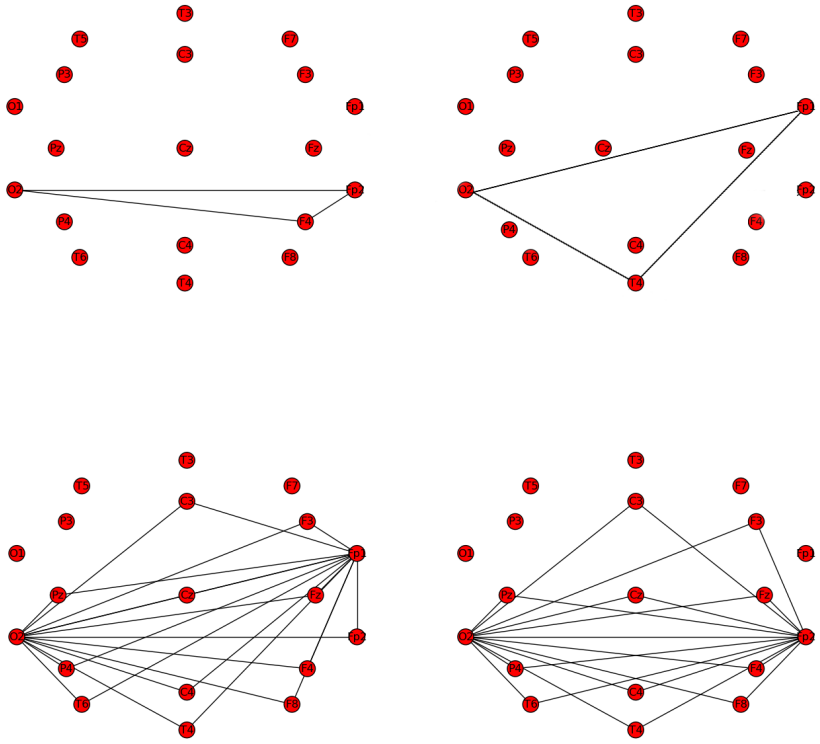


Fig. 3.3. Two of the most significant basic motifs (on the top) and two of the most significant derived motifs (on the bottom) characterizing the tracing segments of patients with MCI from patients with AD

with AD), in Table 3.12, we report some quantitative measures characterizing them. Specifically, the considered measures are: *(i)* the number of edges linking two nodes of the right part of the brain (r-r edges); *(ii)* the number of edges linking a node of the left part and a node of the right part of the brain (l-r edges); *(iii)* the number of edges linking two nodes of the left part of the brain (l-l edges); *(iv)* the number of

edges linking a node of the central part and a node of the right part of the brain (c1-r edges); *(v)* the number of edges linking a node of the central part and a node of the left part of the brain (c1-l edges); *(vi)* the number of edges linking two nodes of the central part of the brain (c1-c1 edges); *(vii)* the number of edges linking two nodes of the frontoparietal part of the brain (f-f edges); *(viii)* the number of edges linking a node of the frontoparietal part and a node of the occipital part of the brain (f-o edges); *(ix)* the number of edges linking two nodes of the occipital part of the brain (o-o edges); *(x)* the number of edges linking a node of the central part and a node of the frontoparietal part of the brain (c2-f edges); *(xi)* the number of edges linking a node of the central part and a node of the occipital part of the brain (c2-o edges); *(xii)* the number of edges linking two nodes of the central part of the brain (c2-c2 edges).

Let us now examine in detail the two derived motifs shown in Figure 3.3. The former is centered on the electrodes $O2$ and $Fp1$, whereas the latter is centered on the electrodes $O2$ and $Fp2$. The analysis of these motifs provides important information about what happens in the brain areas when a patient converts from MCI to AD. In fact, in both cases, the node $O2$ is central. This indicates that the corresponding brain area is very active in patients with MCI and little active (or inactive) in patients with AD. Furthermore, in both cases, it emerges a very intense activity in the right part of the brain in patients with MCI, which reduces or disappears in patients with AD. This could lead to conclude that the conversion from MCI to AD creates deeper damages in the right part of the brain (especially, in the area corresponding to the electrode $O2$) than in the left one.

As a further confirmation of these results, consider the quantitative values reported in Table 3.12. They show that most of the edges connect two nodes of the right part of the brain and that often one node is situated in the frontopolar area and the other resides in the occipital area.

Parameter	*First Derived Motif*	*Second Derived Motif*
r-r edges	7	13
l-r edges	8	4
l-l edges	2	0
c1-r edges	3	6
c1-l edges	3	0
c1-c1 edges	0	0
f-f edges	5	4
o-f edges	6	5
o-o edges	3	3
c2-f edges	4	4
c2-o edges	4	4
c2-c2 edges	0	0

Table 3.12. Quantitative results representing the derived motifs of Figure 3.3

3.3.6 Comparison with other existing approaches

In this section, we compare our approach with the one illustrated in [292]. In our opinion, this comparison is extremely interesting to highlight the potential of our approach because: *(i)* both our approach and the one of [292] use the same metric (i.e., Permutation Disalignment Index) for evaluating the connection degree of brain areas; *(ii)* the authors of [292] showed that their approach is well suited for evaluating the conversion from MCI to AD, and they support their claim by means of comparisons between their approaches and some related ones proposed in the past.

In [292], the authors used boxplots to verify whether a subject with MCI at t_0 converts to AD at t_1 or not. We applied both the approach of [292] and our own to the EEGs of four patients. Two of them suffered from MCI at both t_0 and t_1, whereas two other ones converted from MCI at t_0 to AD at t_1. Clearly, the number of patients we are considering is very small. However, we point out that we do not aim at precisely quantifying how much the performance of our approach is better (or worse) than the one of the approach of [292]. Actually, we simply want to provide the reader with an idea of the way of proceeding of our approach (which implies the need to graphically show the colored networks and the boxplots associated with the EEGs of the patients we are examining) and, possibly, to give a rough comparative estimation of its performance.

In Figure 3.4, we report the boxplots of the four patients into examination. In Table 3.13, we present the values of some parameters helping us to quantify the results shown therein. Analogously, in Figure 3.5, we present the colored networks of the same four patients. In Table 3.14, we show the values of the corresponding conversion coefficient.

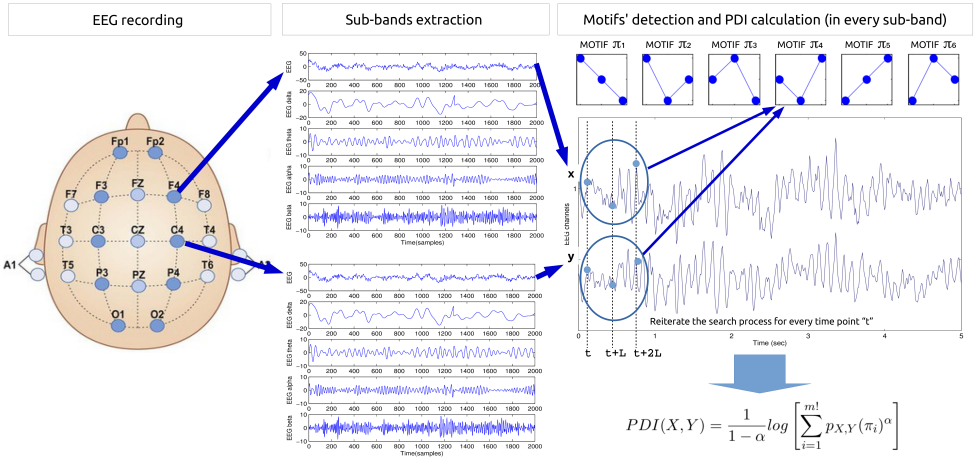


Fig. 3.4. Results of the application of the approach of [292] to the four subjects into consideration

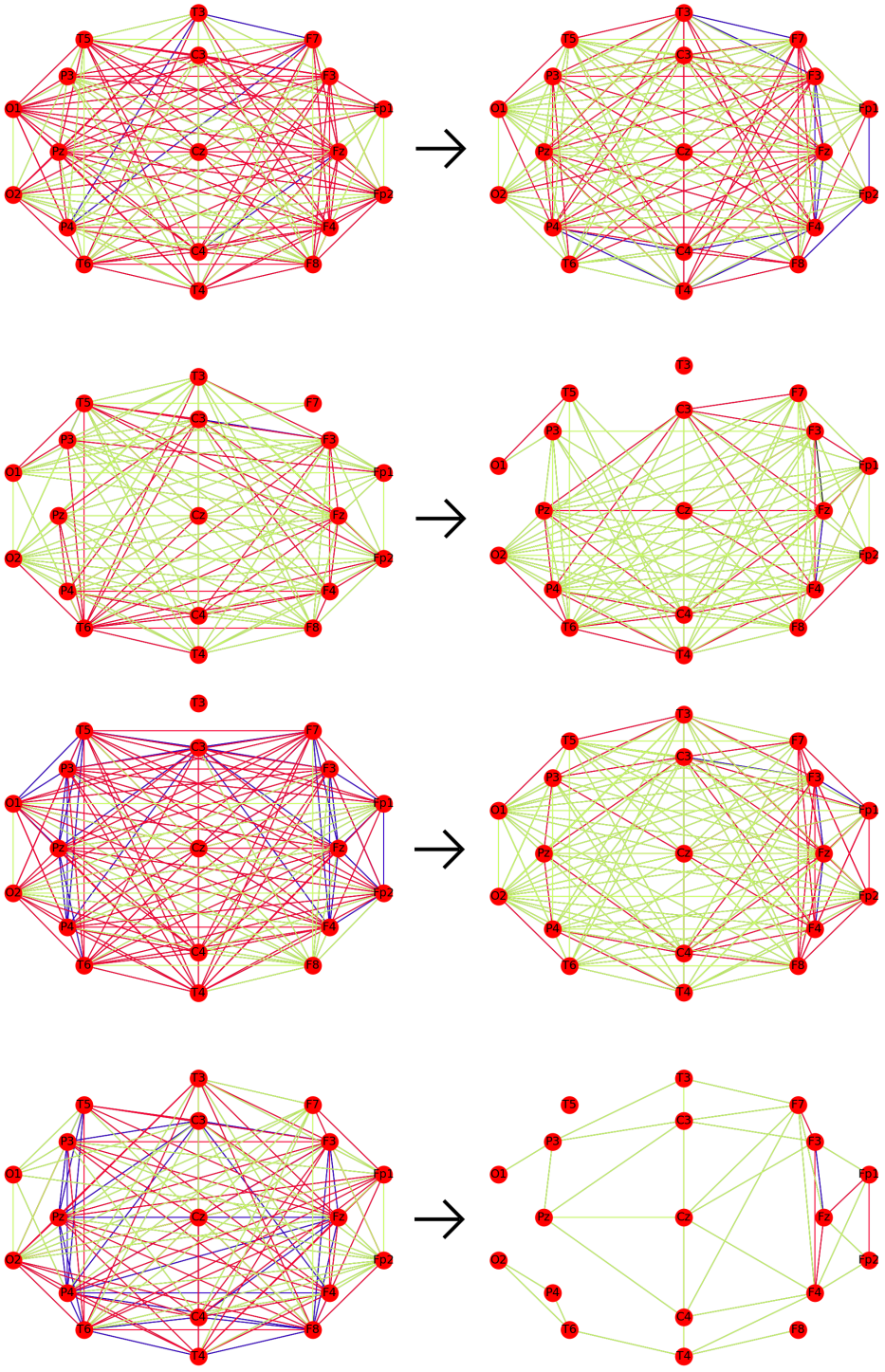


Fig. 3.5. The networks $\mathcal{N}_{0_\pi}$ and $\mathcal{N}_{1_\pi}$ for the two patients not converting to AD (above) and for the two other ones converting to AD (below)

	Variation of medians from $T0$ to $T1$	Variation of 25^{th} percentile from $T0$ to $T1$	Variation of 75^{th} percentile from $T0$ to $T1$
I subject MCI-MCI	9.04%	8.59%	9.13%
II subject MCI-MCI	4.93%	5.75%	4.53%
I subject MCI-AD	20.65%	11.27%	35.97%
II subject MCI-AD	31.70%	19.59%	43.43%

Table 3.13. Quantitative results representing the results shown in Figure 3.4

	Conversion coefficient $conv_{eeg}$
I subject MCI-MCI	-25.00%
II subject MCI-MCI	-4.96%
I subject MCI-AD	-89.06%
II subject MCI-AD	-99.41%

Table 3.14. Values of the conversion coefficient $conv_{eeg}$ for the four patients into examination

From the analysis of Figures 3.4 and 3.5 and from the comparison of Tables 3.13 and 3.14, we can observe that our approach appears more adequate than the one of [292] in distinguishing patients converting from MCI to AD from the ones who do not convert. Indeed:

- When passing from t_0 to t_1 boxplot positions certainly vary more for patients converting to AD than for patients who do not convert. However, this variation is not very clear and marked (see Figure 3.4). Vice versa, when passing from t_0 to t_1, the number and the color of network edges do not present a great variation for patients who do not convert to AD, whereas both these indicators strongly vary for patients converting to AD (see Figure 3.5).
- The variation of medians (resp., 25^{th} percentile and 75^{th} percentile) is about 6.5% (resp., 7%, 6.5%) for patients who do not convert to AD, whereas it is about 26% (resp., 15%, 44%) for patients converting to AD (see Table 3.13).

Instead, if we consider our conversion coefficient, we can observe that its value is about 12% for patients not converting to AD, whereas it is about 9% for patients converting to *AD* (see Table 3.14).

All these evaluations allow us to claim that our approach is really more adequate than the one of [292] to help an expert to visually and quantitatively evaluate the longitudinal history of a patient suffering from MCI and/or AD.

3.3.7 Discussion

Clearly, the results presented in all the previous subsections will require much more efforts and investigations in the future, especially by experts in neurological diseases, in order to completely "capture" their meaningfulness. Nevertheless, they are an interesting "food for thought" that our approach is providing to researchers in this sector.

At the end of this research we can generalize the found results and draw the following hypothesis about the conversion from MCI to AD:

- Conversion coefficient is a well suited indicator of the transition of a patient from MCI to AD. In particular, a decrease of this coefficient of more than 80% in three months is a clear indicator that the corresponding patient is converting from MCI to AD.
- The activity of the brain area underlying the electrode $O2$ and of the right part of the brain is a potential indicator of a possible transition of a patient from MCI to AD. In particular, a marked reduction of the activity of these two brain parts is a possible indicator that the corresponding patients is converting from MCI to AD.

4

Childhood Absence Epilepsy

4.1 Introduction

Nearly 1% of the world population is affected by epilepsy, a neurological disorder characterized by recurrent seizures. Epileptic seizures are still considered unpredictable, despite the huge efforts spent in recent years by scientific community to develop predictive algorithms. These are mainly based on electroencephalography, which consists in recording the scalp potentials produced by cortical electrical activity. Nearly 66% of patients can be successfully treated with anti-epileptic drugs, which have remarkable side effects, whereas nearly 8% of the drug-untractable patients are treated with surgery, which is high-invasive and high-risk. There is no way to treat the remaining 26% of patients.

In this chapter, the attention is focused on Childhood Absence Epilepsy (CAE), an idiopathic generalized epileptic disorder [138, 137] characterized by recurrent "absence seizures" that cause disruption of awareness and are often associated with staring. Subjects experiencing absence seizures must undergo electroencephalography, which is a totally non-invasive and comfortable examination, consisting in recording the cortical electrical activity by means of scalp electrodes that are wired to an acquisition system, connected to a computer.

The electroencephalography acquisition can last from minutes to hours, depending on the number of recorded seizures and of the specific goal of the examination. In order to evaluate electroencephalograms (EEGs), a neurologist manually scrolls them, for detecting and inspecting every possible ictal state (seizure) or abnormality in the interictal (seizure-free) activity. However, manual review is a time-consuming, inefficient and subjective procedure.

To expedite it and to facilitate the diagnosis, worldwide researchers are working to automatically mark the critical events occurring in an EEG, as well as to extract meaningful features from EEG signals, which can help a neurologist to make a diagnosis, to understand the pathology and, therefore, to optimize the treatment.

So far, many methodologies were proposed in the literature for the analysis of EEGs registering absence seizures. Permutation Entropy (PE), a symbolic complexity measure, was introduced in [51] and applied in [88] to analyze epileptic EEGs. Authors of [88] used PE to discriminate the different phases of epileptic activity in intracranial EEG time series, recorded from three intractable patients. In [267], PE was tested as a possible predictor of absence seizures in Genetic Absence Epilepsy Rats from Strasbourg (GAERS). PE outperformed Sample Entropy (SE) and detected the pre-ictal state in 169 out of 314 seizures from 28 rats, and the average anticipation time was $4.9s$. In [81], the authors exploited complexity analysis to detect vigilance changes in epileptic patients. In [343], Multiscale Permutation Entropy (MPE) was proposed to analyze human EEG signals at different absence seizure states. MPE, used in conjunction with Linear Discriminant Analysis (LDA), achieved a 90.6% sensitivity and exhibited a reduction of MPE levels from the inter-ictal state to the ictal one. In [484], the authors proposed Multi-Scale K-means (MSK-means) unsupervised learning to classify epileptic EEG signals and detect epileptic areas. In order to analyze the dynamics of EEG time series, while taking their mutual spatial dependence into account, a spatial-temporal analysis of epileptic EEGs was proposed in [298, 294, 295]. Due to the ability of PE in capturing the dynamics of EEGs registering absence seizures, a PE-based spatial-temporal analysis was proposed in [297, 296, 293]. Here, the authors showed that the frontal temporal lobes exhibited relatively high PE levels, whereas the parieto-occipital areas appeared associated with relatively low PE values. However, being PE univariate, it is only able to quantify the randomness of single EEG channels independently; instead, it is not able to quantify the interaction between channels. To investigate this last issue, the necessity arises of bivariate descriptors, which can provide an estimation of the interaction between channels.

Among this last kind of descriptors, *coherence* is one of the most promising [370]. As a matter of fact, in [384], Partial Directed Coherence (PDC) was employed to quantify the strength and the direction of the interactions between the electrodes during the inter-ictal (i.e., seizure free) EEG segments in CAE patients. PDC revealed an abnormal cortical network activity during the inter-ictal state, in particular in the alpha band. In [377], the authors proposed a method consisting of a three level wavelet decomposition, a coherence estimation and a phase synchrony feature extraction to classify ictal vs inter-ictal EEG segments.

Since absence ictal states appear associated with an increased EEG synchronization, coherence revealed a powerful descriptor of absence seizure EEG signals [384, 370]. In [370], the authors constructed EEG networks, based on the estimation of coherence and Synchronization Likelihood (SL), to investigate the network changes associated with seizure onset. Ictal EEG segments were characterized by an

increased synchronization and a more ordered network topology. In [380], the authors applied PDC-based weighted directed graph analysis to EEGs of patients with absence seizures to perform a classification of nodes (electrodes) according to their source/sink nature.

This chapter aims at providing a contribution in this setting. In fact, it studies the temporal variation of the synchronization between EEG signals to automatically discriminate ictal vs inter-ictal states, while keeping the global view of how these temporal variations involve the different areas of the cortex. For this purpose, an EEG-based complex network model was developed, where nodes represent electrodes (i.e., cortical areas) and the weight of edges represents the complementary of the coherence value between the EEG signals, recorded at the electrodes associated with the corresponding nodes. A complex network-analysis was carried out to find possible changes in the network features driven by the onset of epileptiform activity. By studying the behavior of the network and its subnetworks over time, a global evaluation of the behavior of the cortex is possible, and the presence of seizures can be automatically detected.

The proposed approach differs significantly from previous studies related to EEGs with absence seizure. To our best knowledge, this is the first time that social network analysis is applied to the EEG of patients with absence seizures. Furthermore, in previous studies, based on the use of complex networks for the detection of absence ictal states, for every patient, only one seizure was manually selected [370], whereas, in the present work, all the recorded seizures are considered, and an overall accuracy is achieved for every patient. In our approach, the whole EEG recording is segmented into overlapping windows and, then, it is processed window by window, so that an overall and smooth analysis of it is achieved. Moreover, no artifact rejection preprocessing was carried out in order to introduce no discontinuity in the dataset and to track the dynamics of the EEG time series continuously. Furthermore, a novel complex network parameter, called connection coefficient, is introduced. It proved particularly adequate to quantify the connection level of a network. The present paper is mainly methodological as it introduces a novel approach for the analysis of EEGs with absence seizures. However, since preliminary results, achieved over a dataset of 9 CAE patients, are very encouraging, the proposed method will be tested on a larger dataset in the near future.

This chapter is organized as follows: in Section 4.2, we describe available data and define coherence. In Section 4.3, we introduce some support data structures employed by our approach. Section 4.4 represents the core of this chapter, because it illustrates our approach.

4.2 Available data

4.2.1 EEG recording and preprocessing

A dataset including 9 EEG recordings from patients diagnosed with CAE was studied. The children mean age was 7.44 years, with a standard deviation of 1.67 years. The average duration of EEG recordings was 25.68 mins.

The dataset was provided by $UNEEG^{TM}$ medical A/S (Lynge, Denmark) within a research cooperation agreement. The EEG montage was set according to the international 10/20 system. EEGs were recorded by means of Stellate Harmonie (Stellate Systems, Inc., Montreal, Quebec, Canada) and Cadwell Easy II (Cadwell Laboratories, Inc., Kennewick, WA) systems. EEG traces were reviewed by a board-certified epileptologist, who marked all the paroxysms.

The method flowchart can be described as follows: 1) the n-channels EEG is recorded, band-pass filtered between 0.5 and 32 Hz (because absence seizure activity mainly lies in this range [175]), digitized with a sampling rate of 200 Hz and stored on a computer; 2) the EEG is segmented into M overlapping windows (with $2s$ width and $1s$ overlap) and analyzed window by window; 3) given the k^{th} window *EEG(k)* (where $k = 1, \ldots, M$), the complementary $1 - C_{v_i,v_j}$ of the coherence between every pair of electrodes (v_i, v_j) is estimated, and used as the weight of the edge between the nodes corresponding to v_i and v_j.

The width of the overlapping windows was set at $2s$ because the paroxysms longer than $2s$ are those considered to be clinically relevant. The $1s$ overlap ensures that the EEG is processed smoothly and that there is no abrupt variation in estimated descriptors. EEG processing was implemented and carried out in MATLAB R2016b (The MathWorks, Inc., Natick, MA, USA).

4.2.2 Coherence estimation

The magnitude squared coherence between two signals v_i and v_j depends on the frequency f and is defined as:

$$C_{v_i,v_j}(f) = \frac{|P_{v_i,v_j}(f)|^2}{P_{v_i,v_i}(f)P_{v_j,v_j}(f)}$$

where $P_{v_i,v_i}(f)$ and $P_{v_j,v_j}(f)$ are the Power Spectral Densities (PSD) of v_i and v_j, respectively, whereas $P_{v_i,v_j}(f)$ represents the cross power spectral density between v_i and v_j. Coherence C_{v_i,v_j} is a measure of synchronization between v_i and v_j and is bounded between 0 and 1. In this work, it was estimated using the method of Welch's averaged, modified periodogram [461]. The k^{th} EEG window under analysis, and the estimated values of coherence $C^k_{v_i,v_j}(f)$ for every frequency f, were averaged over the frequencies of the range under consideration (f_L=0.5 Hz - f_U=32 Hz):

$$\overline{C}^k_{v_i,v_j} = \frac{1}{f_U - f_L} \int_{f_L}^{f_U} C^k_{v_i,v_j}(f) df$$

Therefore, for every analyzed window *EEG(k)*, and for every pair of electrodes (v_i, v_j), an average value of coherence $\overline{C}^k_{v_i,v_j}$ is computed.

4.3 Support data structures

Let *eeg* be a generic EEG. Starting from it, a network (that we call *brain network*) $\mathcal{N} = \langle V, E \rangle$ can be defined.

Here, V is the set of nodes of $\mathcal{N}$. Each node $v_i \in V$ corresponds to an electrode. In our EEGs, electrodes were applied by following the 10-20 system and $|V| = 19$.

E is the set of edges of $\mathcal{N}$. Each edge e_{ij} connects nodes v_i and v_j. It can be represented as $e_{ij} = (v_i, v_j, w_{ij})$. Here, w_{ij} is a measure of "distance" between v_i and v_j. It is an indicator of the disconnection level of v_i and v_j. Indeed, each measure representing this feature could be adopted in our model. In the experiments presented in this paper, we employed the complementary of the coherence value between v_i and v_j (i.e., we set $w_{ij} = 1 - C_{v_i,v_j}$).

A preliminary investigation performed in our research consisted of determining the edge weight distribution (averaged on all available patients) in ictal, pre-ictal, post-ictal and inter-ictal states, even if, in this chapter, our focus is on ictal and inter-ictal states. In carrying out this task, we separated the range of edge weights (which, we recall, is $[0, 1]$) in ten intervals of the same length. The obtained distribution for inter-ictal and ictal states is reported in Figures 4.1 and 4.2. From a deeper evaluation of these distributions, we can observe that there are some intervals more relevant than others for distinguishing the two states. As a consequence, to better detect and characterize ictal states, it is reasonable to consider some ad-hoc subnetworks, each taking only the edges belonging to specific intervals into account.

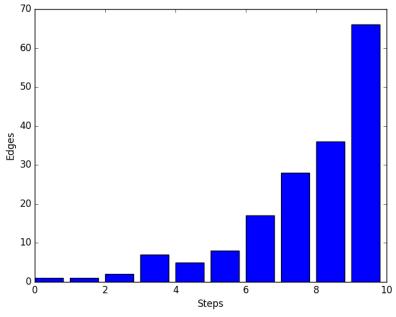


Fig. 4.1. Average edge weight distribution in inter-ictal states

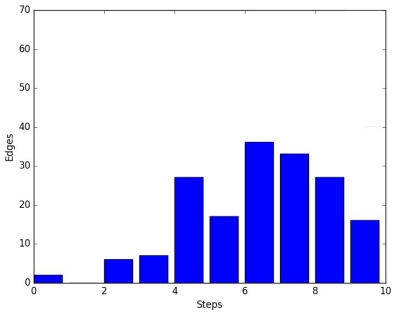


Fig. 4.2. Average edge weight distribution in ictal states

In particular, we defined the following subnetworks: $\mathcal{N}^b = \langle V, E^b \rangle$, $\mathcal{N}^m = \langle V, E^m \rangle$, $\mathcal{N}^r = \langle V, E^r \rangle$, $\mathcal{N}^g = \langle V, E^g \rangle$, $\mathcal{N}^y = \langle V, E^y \rangle$, $\mathcal{N}^{br} = \langle V, E^{br} \rangle$. We used a color name to provide a more mnemonic way of distinguishing sub-networks. Thus, the superscripts b, m, r, g, y, br stand for *blue, magenta, red, green, yellow* and *brown*, respectively.

Given that the set of nodes is the same for all subnetworks, we focus on defining only the set of edges of each of them:

$$E^b = \{e_{ij} \mid e_{ij} \in E, 0.9 < w_{ij} \leq 1\},$$
$$E^m = \{e_{ij} \mid e_{ij} \in E, 0.8 < w_{ij} \leq 0.9\},$$
$$E^r = \{e_{ij} \mid e_{ij} \in E, 0.7 < w_{ij} \leq 0.8\},$$
$$E^g = \{e_{ij} \mid e_{ij} \in E, 0.6 < w_{ij} \leq 0.7\},$$
$$E^y = \{e_{ij} \mid e_{ij} \in E, 0.5 < w_{ij} \leq 0.6\},$$
$$E^{br} = \{e_{ij} \mid e_{ij} \in E, 0.3 < w_{ij} \leq 0.4\}.$$

Finally, we constructed two further subnetworks. The former was obtained by merging *blue*, *magenta*, *red*, *green* and *yellow* subnetworks (we called "*rainbow*" this network). The latter was constructed by considering all the edges of the original network not belonging to the *rainbow* one (we called "*black*" this network). Formally speaking, the two networks are defined as: $\mathcal{N}^{rbw} = \langle V, E^{rbw} \rangle, \mathcal{N}^{blk} = \langle V, E^{blk} \rangle$, where:

$$E^{rbw} = \{e_{ij} \mid e_{ij} \in E, 0.5 < w_{ij} \leq 1\},$$
$$E^{blk} = \{e_{ij} \mid e_{ij} \in E, 0 \leq w_{ij} \leq 0.5\}.$$

Since each EEG is in the form of a time series, it could be useful to introduce the concept of *mean network*. Thus, given q networks $\mathcal{N}_1 = \langle V, E_1 \rangle$, $\mathcal{N}_2 = \langle V, E_2 \rangle$, $\cdots$, $\mathcal{N}_q = \langle V, E_q \rangle$, we define the *mean network* $\overline{\mathcal{N}}$, corresponding to them, as: $\overline{\mathcal{N}} = \langle V, \overline{E} \rangle$, where:

$$\overline{E} = \{(v_i, v_j, \overline{w}_{ij}) \mid e_{ij_k} = (v_i, v_j, w_{ij_k}) \in E_k, 1 \leq k \leq q, \overline{w}_{ij} = \frac{\sum_{k=1}^{q} w_{ij_k}}{q}\}.$$

Observe that $0 \leq \overline{w}_{ij} \leq 1$.

4.4 Detection and characterization of ictal states

4.4.1 Connection coefficient

As pointed out before, one of the main features to investigate for the detection of ictal states is the connection level of the brain areas. In network analysis, one of the most powerful tools for investigating the connection level of a network is the concept of clique [439]. Starting from cliques, it is possible to define a quantitative coefficient, which we call *connection coefficient*, capable of measuring the connectivity level of a network associated with an EEG. This coefficient takes the following considerations into account: *(i)* both the dimension and the number of cliques are important as connectivity indicators; *(ii)* the concept of clique is intrinsically exponential; in other words, a clique of dimension $n+1$ is exponentially more complex than a clique of dimension n. We are now able to define the connection coefficient $cc_{\mathcal{N}}$ of a network $\mathcal{N}$. In particular, let $\mathcal{C}$ be the set of the cliques of $\mathcal{N}$; let $\mathcal{C}_k$ be the set of cliques of dimension k of $\mathcal{N}$; finally, let $|\mathcal{C}_k|$ be the cardinality (i.e., the number of cliques) of $\mathcal{C}_k$. Then, $cc_{\mathcal{N}}$ is defined as:

$$cc_{\mathcal{N}} = \sum_{k=1}^{|V|} |\mathcal{C}_k| \cdot 2^k$$

4.4.2 Detecting ictal states

Detecting ictal states is a very delicate and time consuming task for a neurologist, who have to analyze a whole EEG. Our effort, in this case, was to compute, on a time-slot base, the value of the connection coefficient for an EEG. And so, for each patient and each time-slot, we computed the value of the connection coefficient of the brain network associated with the EEG at that time-slot.

Indeed, in order to better evidence this phenomenon, we considered the brain subnetworks $\mathcal{N}^{rbw}$ and $\mathcal{N}^{blk}$, defined in Section 4.4. We recall that $\mathcal{N}^{rbw}$ considers the five intervals of edge distribution characterized by the heaviest weights, whereas $\mathcal{N}^{blk}$ encompasses the other ones. As a consequence, since edge weights represent distances, on the basis of the results of [370], we can expect that, in presence of an ictal state, the connection coefficient associated with $\mathcal{N}^{rbw}$ presents a minimum, whereas the one corresponding to $\mathcal{N}^{blk}$ shows a maximum. This is explained by the fact that, during the ictal states, the weights of the edges tend to decrease and, therefore, several edges disappear from $\mathcal{N}^{rbw}$ and appear in $\mathcal{N}^{blk}$. For the sake of brevity, we will only look at the results obtained for Patient 18 as an example, but we obtained very similar results for all of the other patients.

In Table 4.1, we report data about figures specified by an expert neurologist when she examined the whole EEG of Patient 18. The physician identified 8 seizures, which took place into the time-slots specified in this table.

Table 4.1. Table produced by a neurologist about start and end time-slots for each seizure of Patient 18

Seizure id	*Start time-slot*	*End time-slot*
1	4	26
2	120	122
3	165	205
4	306	332
5	449	451
6	470	496
7	642	659
8	891	913

We use this table as a starting point and a benchmark of accuracy for the detection of ictal states performed by our approach.

In Figure 4.3, we plotted the values of the connection coefficient (y axis) for each time-slots (x axis) for Patient 18 and for $\mathcal{N}^{rbw}$, whereas, in Figure 4.4, we represented the values of the same coefficient for the same patient, but for $\mathcal{N}^{blk}$.

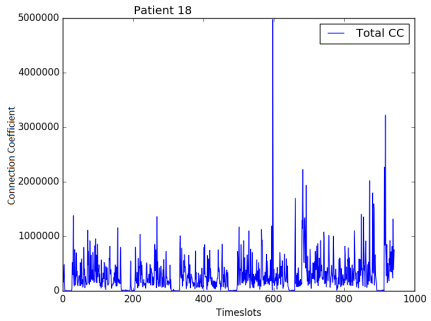


Fig. 4.3. Connection coefficient for the network $\mathcal{N}^{rbw}$ of Patient 18

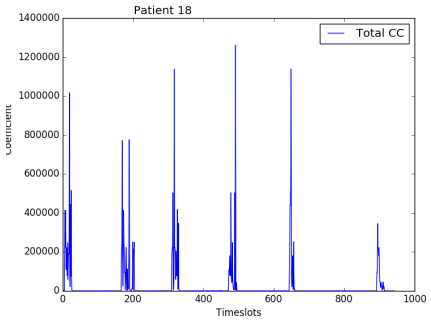


Fig. 4.4. Connection coefficient for the network $\mathcal{N}^{blk}$ of Patient 18

Clearly, in Figure 4.3 it is straightforward to observe that there are some time-slots in which connection coefficient is several orders of magnitude less than others. The important result is that those time-slots are exactly the ones that the neurologist

spotted as ictal states. For instance, in Figures 4.5 and 4.6, we show more closely the part of the plot of Figure 4.3 corresponding to the first and the eighth seizures, in such a way as to allow the reader to more appreciate the differences, in terms of magnitude, of the value of connection coefficient in the involved states.

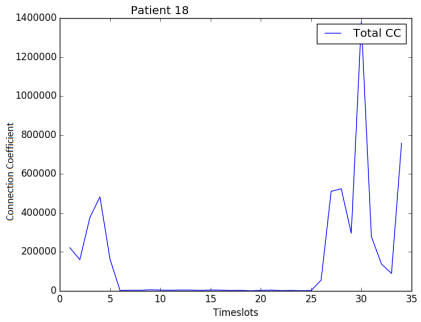


Fig. 4.5. Zoomed plot of the value of connection coefficient of Figure 4.3 - first seizure

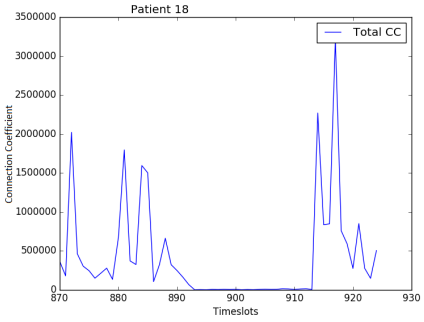


Fig. 4.6. Zoomed plot of the value of connection coefficient of Figure 4.3 - eighth seizure

Thus, without having to manually analyze the whole EEG for a patient, thanks to this coefficient, we can easily distinguish ictal states from the others.

In order to provide a quantitative evaluation of the performance of our approach, we computed its sensitivity, specificity and precision for each patient and, then, for the set of seizures of all patients, taken as a whole. Obtained results are reported in Table 4.2. Taking into account that, in this application context, sensitivity is more important than specificity, we have considered the union of the seizures detected by using $\mathcal{N}^{rbw}$ and $\mathcal{N}^{blk}$.

From the analysis of this table, we can see that our approach provides excellent results, especially if we look at sensitivity. However, also specificity and precision are very good. Clearly, we are conscious that the number of examined patients is small. However, as previously pointed out, due to the encouraging results obtained, and due the *methodological nature* of our paper, we believe the present research can contribute

Table 4.2. Sensitivity, Specificity and Precision of our approach

Patient	*Sensitivity*	*Specificity*	*Precision*
16	0.9032	0.8400	0.7272
18	1.0000	0.9291	0.9961
23	0.9629	0.9882	0.9167
29	1.0000	0.9483	0.9473
31	1.0000	0.9287	0.9438
32	1.0000	0.9356	0.8644
39	1.0000	0.8642	0.7400
47	1.0000	0.9610	0.9917
57	1.0000	0.9012	0.4375
Overall	0.9704	0.9169	0.6482

to motivate clinical centers to engage an experimentation of our approach with a much higher number of patients.

4.4.3 Characterizing ictal states

In order to understand and characterize what happens during *ictal states*, we analyzed the subnetworks defined in Section 4.3. For this purpose, we computed a mean network for each inter-ictal time-slot, up to 40 time slots before the seizures, and we mediated those networks among all patients. We did the same task for ictal time-slots, up to 8 time-slots after the start of a seizure, until the *center-ictal time-slot*.

The subnetworks we used were the blue one $\mathcal{N}^b$, the magenta one $\mathcal{N}^m$, the red one $\mathcal{N}^r$, the green one $\mathcal{N}^g$, the yellow one $\mathcal{N}^y$ and the brown one $\mathcal{N}^{br}$. We computed the values of connection coefficient for each colored subnetwork of the mean networks previously derived. Obtained results are plotted in Figure 4.7. Here, we show the average connection coefficient of some colored networks for 40 time-slots of inter-ictal state and 8 time-slots of ictal state.

As we can see from this figure, during the inter-ictal state, the values of connection coefficient do not deeply change until the first time-slot before *ictal*. At this time, we can see that the values of connection coefficient increase for yellow and green subnetworks and become higher than the corresponding ones of magenta and blue subnetworks. An increase of the values of connection coefficient for yellow and green subnetworks, coupled with a strong decrease of the values of this coefficient for blue and magenta subnetworks, implies that, during ictal states, both the number and the dimension of the cliques in yellow and green subnetworks increase, whereas the corresponding ones in blue and magenta subnetworks decrease. In turn, this implies that a certain number of edges migrate from magenta and blue subnetworks to green

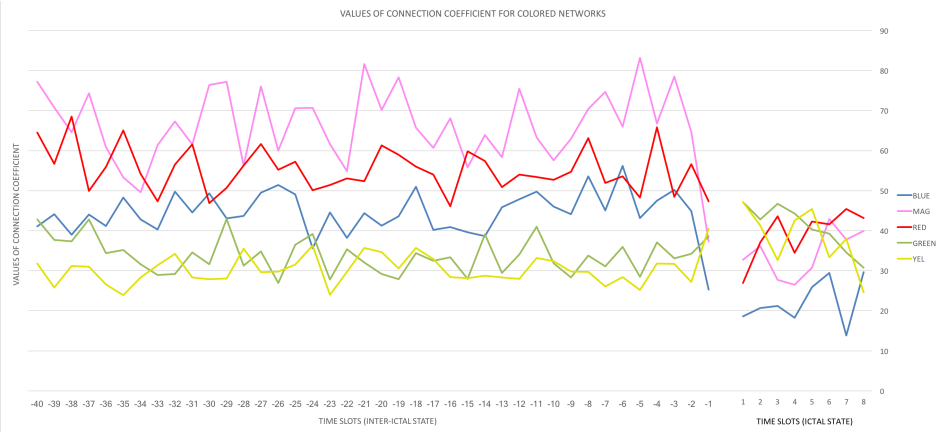


Fig. 4.7. Connection Coefficient for mean networks during pre-ictal and ictal states

and yellow ones. Now, recall that yellow and green edges have a weight between 0.5 and 0.7, whereas magenta and blue edges have a weight between 0.8 and 1. As a consequence, the edge migration described above implies that a hyper-synchronization of brain areas happens during ictal states.

This characterization result for ictal state is particularly interesting because we were able to confirm, through a network analysis-based approach, what several authors had found in the past, through completely different approaches (see, for instance, [370]), namely that ictal states are characterized by hyper-synchronization, which can be automatically detected. With reference to this feature, it is worth emphasizing that we evaluated the sensitivity and the specificity of the proposed approach over the whole EEG recording, and not over selected epochs, which makes our approach suitable for possible real-time applications. Interestingly, in our tests, no artifactual epoch was discarded, in order to track the behavior of the EEG continuously and to evaluate the sensitivity, specificity and precision of our approach in real conditions, when noise and artifacts may be present. Furthermore, the usage of complex networks allows the investigation of the interactions between the different areas of the brain in absence and in presence of a seizure, which we aim at deepening in the future. Finally, we point out that, at the moment, the system can be used off-line to mark the seizures automatically and allow the neurologist to skip the manual EEG review, which is extremely time consuming. However, we plan to optimize it in the future in such a way as to allow for a continuous, real time, long-term monitoring.

Part II

Data Lakes

In this part, we apply our network-based model and the associated social network-based approach to data lake management. In particular we propose a new metadata model well suited for data lakes. Our model starts from the considerations and the ideas proposed by data lake companies (in particular, it starts from the general metadata classification also used by Zaloni [341]). However, it complements them with new ideas and, in particular, with the power guaranteed by a network-based and semantics-driven representation of metadata.

This part is organized as follows: in Chapter 5, we present an approach to uniformly handle heterogeneous Data Lake sources. In Chapter 6, we illustrate our approach for the extraction of interschema properties. Finally, in Chapter 7, we present an approach to the extraction of complex knowledge patterns among concepts belonging to different sources.

5

Uniform Management of Heterogeneous Data Lake Sources

5.1 Introduction

Metadata have always played a key role in favoring the cooperation of heterogeneous data sources [124, 60, 373] [345]. This role was already relevant in the past architectures (e.g., Cooperative Information Systems and Data Warehouses) but has become much more crucial with the advent of data lakes [148]. Indeed, in this new architecture, metadata represent the only possibility to guarantee an effective and efficient management of data source interoperability. As a proof of this, the main data lake companies are performing several efforts in this direction (see, for instance, the metadata organization proposed by Zaloni, one of the market leaders in the data lake field [341]). For this reason, the definition of new models and paradigms for metadata representation and management represents an open problem in the data lake research field.

In this chapter, we aim at providing a contribution in this setting and we propose a new metadata model well suited for data lakes. Our model starts from the considerations and the ideas proposed by data lake companies (in particular, it starts from the general metadata classification also used by Zaloni [341]). However, it complements them with new ideas and, in particular, with the power guaranteed by a network-based and semantics-driven representation of metadata. Thanks to this choice, our model can benefit from all the results already found in network theory and semantics-driven approaches. As a consequence, it can allow a large variety of sophisticated tasks that the metadata models currently adopted do not guarantee. For instance, it allows the definition of a structure for unstructured data, which currently represent more than 80% of available data sources. Furthermore, it allows the extraction of thematic views from data sources [44], i.e., the construction of views concerning one or more topics of interest for the user, obtained by extracting and merging data coming from different sources. This problem has been largely investigated in the past for structured and semi-structured data sources stored in a data warehouse, and this witnesses its

extreme relevance. These are only two of the tasks that can benefit from our model and, in this chapter, we illustrate them. Actually, many other ones could be thought and investigated, and they will represent the subject of our future research efforts.

This chapter is structured as follows: Section 5.2 illustrates related literature. In Section 5.3, we propose our metadata model. Section 5.4 presents the application of this model to the problems of structuring unstructured data and of extracting thematic views from heterogeneous data lake sources. In Section 5.5, we present our example case.

5.2 Related Literature

In the literature, several metadata classifications have been proposed in the past. For instance, the authors of [65] propose a tree-based classification. They split metadata into several categories, propose a conceptual schema of the metadata repository and use RDF for metadata modeling. The strength of this model is undoubtedly its richness, whereas its weakness is its complexity that cannot guarantee a fast processing of the corresponding data.

A metadata model well suited for data lakes in proposed in [341]. This is also the model adopted by Zaloni. It divides metadata based on their generation time or on the meaning and information they bring. In this latter case, metadata can be divided in three categories, namely operational, technical and business metadata. As will be clear in the following, our metadata model starts from this, but it goes much further. In particular, it assumes that the three classes are not independent from each other because there are several intersections of them. Some of these intersections are particularly expressive and important; for them, it provides a network-based representation rich enough to allow several interesting tasks, but, at the same time, not excessively complex in such a way as to prevent a slow processing.

Several metadata models and frameworks are widely adopted by the Linked Data community (e.g., DCMI Metadata Terms and VoID). DCMI Metadata Terms is a set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative. It includes generic metadata, represented as RDF properties, on dataset creation, access, data provenance, structure and format. A subset was also published as ANSI/NISO and ISO standards and as IETC RFC. The Vocabulary of Interlinked Datasets (VoID) is an RDF Schema vocabulary that provides terms and patterns for describing RDF datasets. It is intended as a bridge between the publishers and the users of RDF data. It focuses on: *(i) general metadata*, following the Dublin Core model; *(ii) access metadata*, describing how RDF data can be accessed by means

of several protocols; *(iii) structural metadata*, describing the structure and the schema of datasets, mostly used for supporting querying and data integration.

As for the applications of our metadata model proposed in this chapter (i.e., structuring of unstructured data and thematic view extraction), most approaches proposed in the literature to carry out this task do not completely fit the data lake paradigm. Two surveys on this issue can be found in [189, 12].

Another family of approaches leverages materialized views to perform tree pattern querying [452] and graph pattern queries [147]. Unfortunately, all these approaches are well-suited for structured and semi-structured data, whereas they are not scalable and lightweight enough to be used in a dynamic context or with unstructured data. Interesting advances in this area can be found in [412, 67, 44].

Finally, semantic-based approaches have long been used to drive data integration in databases and data warehouses. More recently, in the context of big data, formal semantics has been specifically exploited to address issues concerning data variety/heterogeneity, data inconsistency and data quality in such a way as to increase understandability. In the data lake scenario, semantic techniques have been successfully applied to more efficiently integrate and handle both structured and unstructured data sources by aligning data silos and better managing evolving data model (see, for instance, [188, 149]). Similarly to what happens in our approach, knowledge graphs in RDF are used to drive integration. To reach their objectives, these techniques usually rely on tools assisting users in linking metadata to uniform vocabularies (e.g., ontologies or knowledge repositories, such as DBpedia).

5.3 A unifying model for representing the metadata of data lake sources

In this section, we illustrate our network-based model to represent and handle the metadata of a data lake, which we will use in the rest of this chapter.

Our model represents a data lake DL as a set of m data sources: $DL = \{D_1, D_2, \cdots, D_m\}$. A data source $D_k \in DL$ is provided with a rich set $\mathcal{M}_k$ of metadata. We denote with $\mathcal{M}_{DL}$ the repository of the metadata of all the data sources of DL: $\mathcal{M}_{DL} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$.

5.3.1 Typologies of metadata

Following what it is said in [341], metadata can be divided into three categories, namely: *(i) Business metadata*, which include business rules (e.g., the upper and lower limit of a particular field, integrity constraints, etc.); *(ii) Operational metadata*,

which include information generated automatically during data processing (e.g., data quality, data provenance, executed jobs); *(iii) Technical metadata*, which include information about data format and schema. Based on this reasoning, $\mathcal{M}_k$ can be represented as the union of three sets $\mathcal{M}_k^B \cup \mathcal{M}_k^O \cup \mathcal{M}_k^T$.

As an advancement of the model of [341], we observe that these three subsets are intersected with each other (as shown in Figure 5.1). For instance, since business metadata contain all business rules and information allowing to better understand data fields, and since the data schema is included in the technical metadata, we can conclude that data fields represent the perfect intersection between these two subsets. Analogously, technical metadata contain the data type and length, the possibility that a field can be NULL or auto-incrementing, the number of records, the data format and some dump information. These last three things are in common with operational metadata, which contain information like sources and target location and the file size as well. Finally, the intersection between operational and business metadata represents information about the dataset license, the hosting server and so forth (e.g. see the DCMI Metadata Terms).

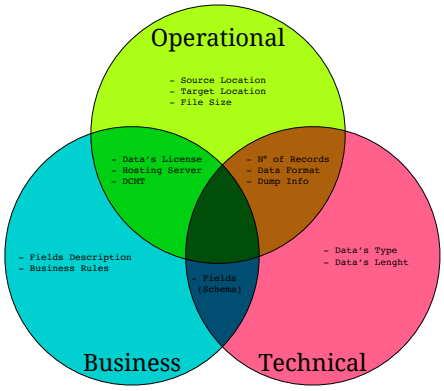


Fig. 5.1. The three kinds of metadata proposed by our model.

In this chapter, we focus on business metadata and on the intersection between them and the technical ones. This intersection contains the data fields, both domain description and technical details. For instance, in a structured database, this intersection contains the attributes of the tables. Instead, in a semi-structured one, it consists of the names of the (complex or simple) elements and attributes of the schema. Finally, in an unstructured source, it could consist of a set of keywords generally adopted to give an idea of the source content.

5.3.2 A network-based model for business and technical metadata

As already mentioned, in this chapter we focus especially on the business and technical metadata and on their intersection. Indeed, they denote, at the intensional level, the information content stored in the data lake sources and are those of interest for supporting most tasks, including the ones described in this chapter.

We indicate by $\mathcal{M}_k^{BT}$ the intersection between $\mathcal{M}_k^B$ and $\mathcal{M}_k^T$. We denote by Obj_k the set of all the objects stored in $\mathcal{M}_k^{BT}$. The concept of "object" depends on data source typology. For instance, in a relational database, objects denote its tables and their attributes. In an XML document or in a JSON one, objects include complex/simple elements and their attributes.

In order to represent $\mathcal{M}_k^{BT}$, our model relies on a suitable directed graph G_k^{BT} $= \langle N_k, A_k \rangle$. For each object $o_{k_j} \in Obj_k$ there exists a node $n_{k_j} \in N_k$. As there is a one-to-one correspondence between a node of N_k and an object of Obj_k, in the following, we will use the two terms interchangeably.

On the other hand, each $a_{k_i} = \langle (n_s, n_t), l_{k_i} \rangle \in A_k$ is an arc; here, n_s is the source node, n_t is the target one, whereas l_{k_i} is a label representing the kind of relationhip between n_s and n_t. Some possible relationships are: *(i) Structural relationship*: it is represented by the label "contains" and is used to represent the relationhip between a relational table and its attributes, a complex object and its simple ones, or between a simple object and its attributes. *(ii) Similarity relationship*: it is represented by the label "similarTo" and denotes a form of similarity between two objects. We will see an example of its semantics and usage in Section 5.4.1. *(iii) Lemma relationship*: it is represented by the label "lemma" and denotes that the target node is a lemma of the source one. Again, its usage will be clear in Section 5.4.1.

Our model enables a scalable and flexible approach in the representation and management of metadata of heterogeneous data lake sources. Indeed, adding a new data source only requires the extraction of its metadata and their conversion to our model. Furthermore, the integration of metadata regarding different data sources can be simply performed by adding suitable arcs between the nodes for which there exists some relationship.

Similarly, G_k^{BT} can be extended with external knowledge graphs (e.g., DBpedia[1]). In the following, we refer to an extension of G_k^{BT} as G_k^{Ext}. It consists of $G_k^{Ext} = G_k^{BT} \cup G^E$, where G^E is an external knowledge graph. An arc from a node of G_k^{BT} and its corresponding node in G^E will be labeled as "externalSource_X", where X is the name of the external knowledge graph at hand.

[1] `http://wiki.dbpedia.org`

5.4 Examples of applications of our metadata model

As pointed out in the Introduction, in order to give an idea of the expressiveness and the power of our data model, in this section, we will exploit it in two application tasks, namely "structuring" unstructured data sources and extracting thematic views from heterogeneous data lake sources.

5.4.1 Defining a structure for unstructured sources

Based on a generic graph representation, our model is perfectly fitted for representing and managing both structured and semi-structured data sources. The highest difficulty regards unstructured data because it is worth avoiding a flat representation, consisting of a simple element for each keyword provided to denote the source content. As a matter of fact, this kind of representation would make the reconciliation, and the next integration, of an unstructured source with the other (semi-structured and structured) ones of the data lake very difficult. Therefore, it is necessary to (at least partially) "structure" unstructured data. Our approach to addressing this issue consists of four phases.

During the first phase, it creates a node representing the source as a whole and a node for each keyword. Then, it links the former to the latter through arcs with label "contains". During the second phase, it adds an arc with label "lemma" from the node n_{k_1}, corresponding to the keyword k_1, to the node n_{k_2}, corresponding to the keyword k_2, if k_2 is registered as a lemma[2] of k_1 in a suitable thesaurus (we adopted BabelNet [326] for this purpose). During the third phase, our approach derives lexical similarities. In particular, it states that there exists a similarity between the nodes n_{k_1}, corresponding to the keyword k_1, and n_{k_2}, corresponding to the keyword k_2, if k_1 and k_2 have at least one common lemma in a suitable thesaurus. Also in this case, we have adopted BabelNet. After having found lexical similarities, it derives string similarities and states that there exists a similarity between n_{k_1} and n_{k_2} if the string similarity degree $kd(k_1, k_2)$, computed by applying a suitable string metric on k_1 and k_2, is higher than a suitable threshold th_k. After several experiments, we have chosen N-Grams [241] as string similarity metric. In both these cases, if there exist a similarity between n_{k_1} and n_{k_2}, our approach adds an arc with label "similarTo" from n_{k_1} to n_{k_2}, and vice versa. During the fourth phase, if there exists a pair of arcs with label "similarTo" between two nodes n_{k_i} and n_{k_j}, our approach merges them into one node $n_{k_{ij}}$, which inherits all the incoming and outgoing edges of n_{k_i} and n_{k_j}.

[2] In this chapter, we use the term "lemma" according to the meaning it has in BabelNet [326]. Here, given a term, its lemmas are other objects (terms, emoticons, etc.) contributing to specify its meaning.

Finally, if there exist two or more arcs from a node n_{k_i} to a node n_{k_j} with the same label, our approach merges them into one node[3].

5.4.2 An approach to extracting thematic views

Our approach to extracting thematic views operates on a data lake DL whose data sources are represented by means of the model described in Section 5.3. It consists of two steps, the former mainly based on the structure of the sources at hand, the latter mainly focusing on the corresponding semantics.

Step 1 of our approach receives a data lake DL, a set of topics $T = \{T_1, T_2, \cdots, T_l\}$, representing the themes of interest for the user, and a dictionary Syn of synonymies involving the objects stored in the sources of DL. This dictionary could be a generic thesaurus, such as BabelNet [326], a domain-specific thesaurus, or a dictionary obtained by taking into account the structure and the semantics of the sources, which the corresponding objects refer to (such as the dictionaries produced by XIKE [124], MOMIS [60] or Cupid [288]). Let T_i be a topic of T. Let $Obj_i = \{o_{i_1}, o_{i_2}, \cdots, o_{i_q}\}$ be the set of the objects synonymous of T_i in DL. Let $N_i = \{n_{i_1}, n_{i_2}, \cdots, n_{i_q}\}$ be the corresponding nodes. First, our approach constructs the ego networks $E_{i_1}, E_{i_2}, \cdots, E_{i_q}$ having $n_{i_1}, n_{i_2}, \cdots, n_{i_q}$ as the corresponding egos. Then, it merges all the egos into a unique node n_i. In this way, it obtains a unique ego network E_i from $E_{i_1}, E_{i_2}, \cdots, E_{i_q}$. If a synonymy exists between two alters belonging to different ego networks, then these are merged into a unique node and the corresponding arcs linking them to the ego n_i are merged into a unique arc. At the end of this task, we have a unique ego network E_i corresponding to T_i. After having performed the previous task for each topic of T, we have a set $E = \{E_1, E_2, \cdots, E_l\}$ of l ego networks. At this point, Step 1 finds all the synonymies of Syn involving objects of the ego networks of E and merges the corresponding nodes. After all the possible synonymies involving objects of the ego network of E have been considered and the corresponding nodes have been merged, a set $V = \{V_1, \cdots, V_g\}$, $1 \leq g \leq l$, of networks representing potential views is obtained. If $g = 1$, then there exists a unique thematic view comprising all the topics required by the user. Otherwise, there exist more views each comprising some (but not all) of the topics of interest for the user.

Step 2 starts by constructing the graph G_k^{Ext} obtained by extending G_k^{BT} with an external knowledge graph G^E (in this work, we rely on DBpedia). For this purpose, first it links each node n_{i_j} of V_i to the corresponding entry $n_{e_{ij}} \in G^E$ through an arc with label "externalSource_DBpedia". In our scenario, such a DBpedia node $n_{e_{ij}}$ is

[3] Please note that Phases 3 and 4 could be merged in a unique one, avoiding to define arcs with label "similarTo". Here, we maintain these arcs and both phases to keep the information about similarity between nodes for future use.

already specified in the BabelNet entry corresponding to n_{i_j} (or to any of its synonyms in Syn)[4]. Then, for each $n_{e_{ij}}$ considered above, all the related concepts are retrieved. In DBpedia, knowledge is structured according to the Linked Data principles, i.e. as an RDF graph built by triples. Each triple $\langle s(ubject), p(roperty), o(bject)\rangle$ states that a subject s has a property p, whose value is an object o. Therefore, retrieving the related concepts for a given element x implies finding all the triples where x is either the subject or the object. For each view $V_i \in V$, the procedure to extend it consists of the following three substeps: *(1) Mapping*: for each node $n_{i_j} \in V_i$, its corresponding DBpedia entry $n_{e_{ij}}$ is found. *(2) Triple extraction*: all the related triples $\langle n_{e_{ij}}, p, o\rangle$ and $\langle s, p, n_{e_{ij}}\rangle$, i.e., all the triples in which $n_{e_{ij}}$ is either the subject or the object, are retrieved. *(3) View extension*: for each retrieved triple $\langle n_{e_{ij}}, p, o\rangle$ (resp., $\langle s, p, n_{e_{ij}}\rangle$), V_i is extended by defining a node for the object o (resp., s), if not already existing, linked to n_{i_j} through an edge labeled as p. Substeps 2 and 3 are recursively repeated for each new added node. The procedure stops after a given number of iterations, limiting the length of external incoming and outcoming paths of nodes in V_i. The longer the path, the weaker the semantic link between nodes.

The enrichment procedure is performed for all the views of V. It is particularly important if $|V| > 1$ because the new derived relationships could help to merge the thematic views that was not possible to merge during the Step 1. In particular, let $V_i \in V$ and $V_l \in V$ be two views of V, and let V_i' and V_l' be the extended views corresponding to them. If there exist two nodes $n_{i_h} \in V_i'$ ad $n_{l_k} \in V_l'$ such that $n_{i_h} = n_{l_k}$[5], then they can be merged in one node; in this way, V_i' and V_l' become connected. After all equal nodes of the views of V have been merged, all the views of V could be either merged in one view or not. In the former case, the process terminates with success. Otherwise, it is possible to conclude that no thematic views comprising all the topics specified by the user can be found. In this last case, our approach still returns the enriched views of V and leaves the user the choice to accept of reject them.

5.5 An example case

In this section, we present an example case aiming at illustrating the various tasks of our approach. Here, we consider: *(i)* a structured source, called *Weather Conditions* (W, in short), whose corresponding E/R schema is not reported for space limitations; *(ii)* two semi-structured sources, called *Climate* (C, in short) and *Environment*

[4] Whenever this does not happen, the mapping can be automatically provided by the DBpedia Lookup Service (`http://wiki.dbpedia.org/projects/dbpedia-lookup`).

[5] Here, two nodes are equal if the corresponding name coincide.

(E, in short), whose corresponding XML Schemas are not reported for space limitations; *(iii)* an unstructured source, called *Environment Video* (V, in short), consisting of a YouTube video and whose corresponding keywords are: *garden*, *flower*, *rain*, *save*, *earth*, *tips*, *recycle*, *aurora*, *planet*, *garbage*, *pollution*, *region*, *life*, *plastic*, *metropolis*, *environment*, *nature*, *wave*, *eco*, *weather*, *simple*, *fineparticle*, *climate*, *ocean*, *environmentawareness*, *educational*, *reduce*, *power*, *bike*.

By applying the approach mentioned in Section 5.4.2, we obtain the corresponding representations in our network-based model, shown in Figure 5.2[6].

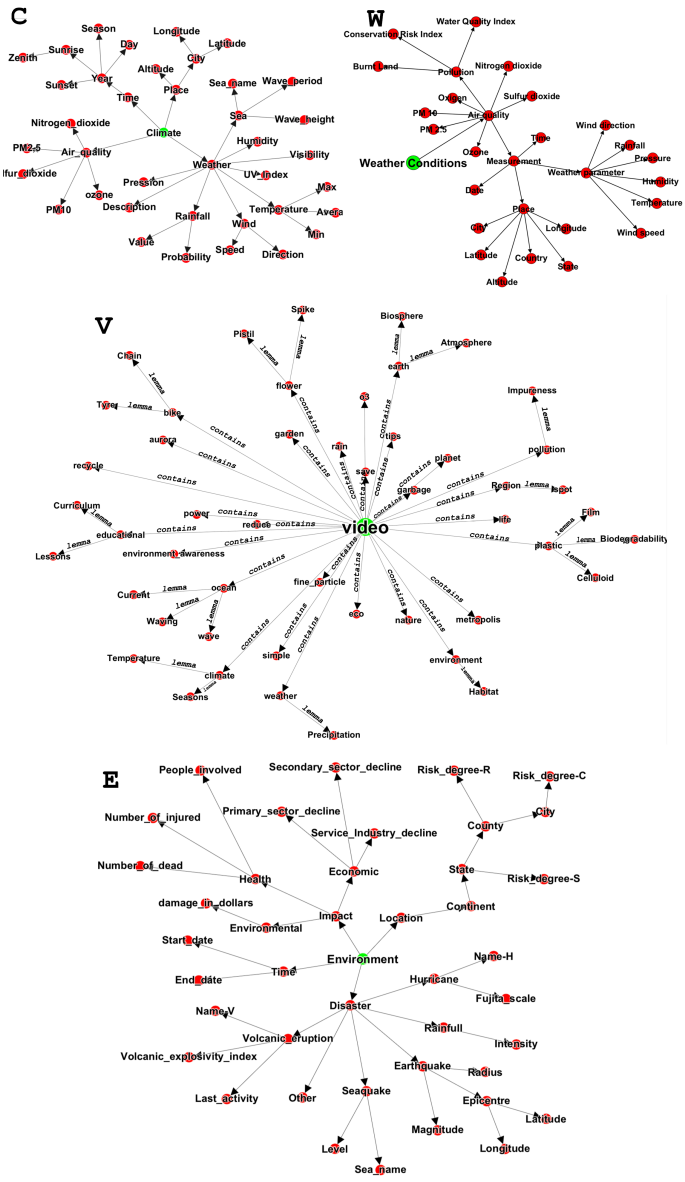


Fig. 5.2. Network-based representations of the four sources into consideration.

[6] In this figure, we do not show the arc labels for the sources C, W and E because all of them are "contains" and their presence would have complicated the layout unnecessarily.

Assume, now, that a user specifies the following set T of topics of her interest: $T = \{Ocean, Area\}$. First, our approach determines the terms (and, then, the objects) in the five sources that are synonyms of *Ocean* and *Area*. As for *Ocean*, the only synonym present in the sources is *Sea*; as a consequence, Obj_1 comprises the node *Ocean* of the source V ($V.Ocean$[7]) and the node *Sea* of the source C ($C.Sea$). An analogous activity is performed for *Area*. At the end of this task we have that $Obj_1 = \{V.Ocean, C.Sea\}$ and $Obj_2 = \{W.Place, C.Place, V.Region, E.Location\}$.

Step 1 of our approach proceeds by constructing the ego networks corresponding to the objects of Obj_1 and Obj_2. They are reported in Figure 5.3[8].

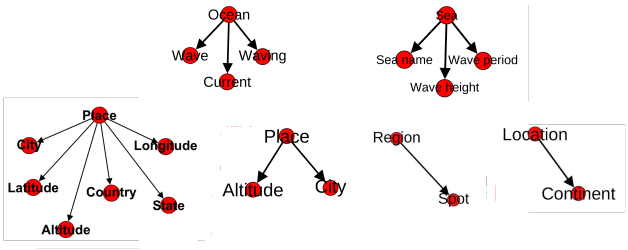


Fig. 5.3. Ego networks corresponding to *V.Ocean*, *C.Sea*, *W.Place*, *C.Place*, *V.Region* and *E.Location*.

Now, consider the ego networks corresponding to *V.Ocean* and *C.Sea*. Our approach merges the two egos into a unique node. Then, it verifies whether further synonyms exist between the alters. Since none of these synonyms exists, it returns the ego network shown in Figure 5.4(a). The same task is performed to the ego networks corresponding to *W.Place*, *C.Place*, *V.Region* and *E.Location*. In particular, first the four egos are merged. Then, synonyms between the alters *W.City* and *C.City* and the alters *W.Altitude* and *C.Altitude* are retrieved. Based on this, *W.City* and *C.City* are merged in one node, *W.Altitude* and *C.Altitude* in another node, the arcs linking the ego to *W.City* and *C.City* are merged in one arc and the ones linking the ego to *W.Altitude* and *C.Altitude* in another arc. In this way, the ego network shown in Figure 5.4(b) is returned. At this point, there are two ego networks, E_{Ocean} and E_{Area}, each corresponding to one of the terms specified by the user.

Step 1 verifies if there are any synonyms between a node of E_{Ocean} and a node of E_{Area}. Since this does not happen, it returns the set $V = \{V_{Ocean}, V_{Area}\}$, where V_{Ocean} (resp., V_{Area}) coincides with E_{Ocean} (resp., E_{Area}).

At this point, Step 2 is executed. As shown in Figure 5.5, first each term (synonyms included) is semantically aligned to the corresponding DBpedia entry (e.g., *Ocean*

[7] Hereafter, we use the notation $S.o$ to indicate the object o of the source S.

[8] In this figure, for layout reasons, we do not show the arc labels because they are the same as the corresponding arcs of Figure 5.2.

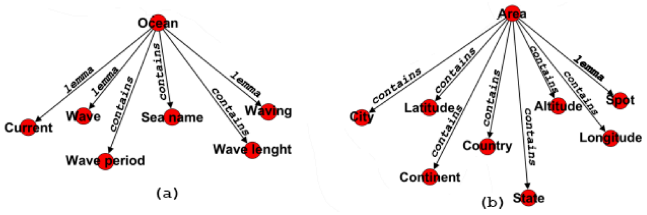


Fig. 5.4. Ego networks corresponding to *Ocean* and *Area*.

is linked to *dbo:Sea*, *Area* is linked to *dbo:Location* and *dbo:Place*, while *Country* to *dbo:Country*[9], respectively). After a single iteration, the following triples are retrieved: ⟨*dbo:sea rdfs:range dbo:Sea*⟩ and ⟨*dbo:sea rdfs:domain dbo:Place*⟩. Other connections can be found by moving to specific instances of the mentioned resources. Indeed, the following triples are retrieved: ⟨*instance rdf:type dbo:Sea*⟩, ⟨*instance rdf:type dbo:Location*⟩, ⟨*instance rdf:type dbo:Place*⟩. Furthermore, a triple ⟨*instance dbo:country dbo:Country*⟩ can be retrieved. As a result, Step 2 succeeded in merging the two views that were separated after Step 1.

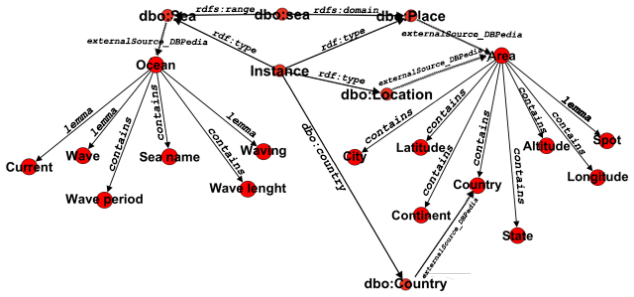


Fig. 5.5. The integrated thematic view.

[9] Prefixes *dbo* and *dbr* stand for `http://dbpedia.org/ontology/` and `http://dbpedia.org/resource/`

6

Extraction of Interschema Properties

6.1 Introduction

In the last few years, we are assisting to a real revolution in the information system scenario. In fact, the number and the size of available data sources have dramatically increased. Furthermore, most of them (i.e., more than 80%) are unstructured [110, 95]. These facts are rapidly changing the scientific and technological "coordinates" of the information system research field [61]. As a consequence of this phenomenon, even issues successfully addressed in the past must be re-considered and re-investigated. One of these issues is certainly the derivation of interschema properties (i.e., *intensional* relationships between concepts represented in different data sources [347], like synonymies, homonymies, hyponymies, overlappings, subschema similarities). This topic has been widely studied in the past [373, 62]; however, the proposed approaches generally considered structured or, at most, semi-structured sources. Furthermore, the number of involved sources, for which most of past approaches were targeted to, was very small, if compared with a typical current source interaction and cooperation scenario.

Interschema property derivation is not just one of the many topics to re-investigate in information systems cooperation field. Actually, it represents the basis of most of the other issues: for instance, the knowledge of interschema properties is necessary for source integration, the construction of data warehouses and data lakes, data analytics, and so forth.

In this chapter, we aim at providing a contribution in this setting. Indeed, we propose a novel approach to uniformly perform the extraction of interschema properties from structured, semi-structured and unstructured sources. Our approach has been specifically conceived having in mind two peculiarities that should characterize it, namely: *(i)* the capability of handling unstructured sources; *(ii)* the lightweightness, making it capable of managing a huge number of data sources.

As for the capability of handling unstructured sources, our approach is provided with a preliminary step capable of "structuring" unstructured sources, i.e., of (at least partially) deriving a structure for them. This is possible because it assumes that each unstructured source (e.g., a video, an audio, an image, a text) has associated a list of keywords describing it. The "structuring" process is based exactly on these keywords. This is another main contribution of this chapter, which, generally speaking, allows the unstructured sources to be uniformly handled as the structured and the semi-structured ones. With regard to this aspect, some clarifications of what we intend with the terms "structured" and "semi-structured" sources are in order. In particular, we use these terms as they are generally adopted in databases and information systems research field. Here, a structured source consists of some concepts, each having a precise set of attributes and relationships with other concepts of the source. A semi-structured source has similar characteristics, but the set of attributes and relationships characterizing a given concept is handled in a more flexible fashion. Indeed, given a property p or a relationship r of a concept c, some instances of c might have exactly one instance of r and/or one instance of p; other instances of c might have more instances of r and/or more instances of p; finally, other ones might have no instances of r and/or no instances of c. A classical example of structured sources is a relational database (that can be conceptually represented by means of an E/R diagram). A classical example of a semi-structured source is an XML document (that can be conceptually represented by means of a DOM).

Unstructured sources are videos, audios, images or texts. They do not generally have a conceptual representation showing their concepts, along with the corresponding properties and relationships. However, they are generally provided with a set of keywords, denoting the main concepts they are representing. The purpose of our approach for "structuring" unstructured sources is exactly the derivation of the relationships existing among the concepts represented by the keywords associated with unstructured sources. If we are capable of performing this task, unstructured sources can be handled similarly to structured and semi-structured ones. Furthermore, their analysis and management could benefit from the wide amount of results found in the past for structured and semi-structured sources. Finally, the integration, the cooperation and the simultaneous querying of structured, semi-structured and unstructured sources are possible.

Our approach also differs from other ones previously presented in related research fields and that could be in principle extended to address the problem we are considering in this chapter. Think, for instance, of ontologies. We could link each available keyword to an ontology and use this last one as the "infrastructure" through which establishing the relationships among the keywords, once these last have been linked to

it. This approach is certainly valid, but it needs a support ontology. As a consequence, it can be employed only in those application fields for which an ontology exists and only if all the involved information sources can be mapped onto a unique ontology. If only some of the involved unstructured sources can be referred to an ontology and/or some of them can be mapped onto another ontology and/or, finally, some of them cannot be referred to any ontology, this way of proceeding cannot be adopted. From this point of view, our approach is more general because it can be applied in all cases, independently of the presence of none, one or more ontologies, which the unstructured sources can be referred to. It only needs a thesaurus. If there exists a specific thesaurus for the scenario which the unstructured sources into examination belongs to, then it uses this theasurus. Otherwise, it can still work with a general-purpose thesaurus, like BabelNet [326]. Clearly, if the unstructured sources are specific of a certain field, the availability of a specific thesaurus can help to obtain a better accuracy. However, if this kind of thesaurus is not available, a general-purpose one is sufficient to proceed even if, in this case, accuracy could be lower.

As for the lightweightness of our approach, we observe that, in a big data scenario, such as the one currently characterizing the information system field, a new proposed approach must take scalability into a primary consideration [270, 268]. As a matter of fact, the sources interacting in every task are always very numerous and large (think, for instance, of a data lake constructed to support data analytics in an organization) and the time allowed for each transaction is very limited (think, for instance, of streaming applications). As a consequence, even approaches considered very scalable in the past (such as DIKE [349], MOMIS [59], and Cupid [288]) are not adequate anymore. In our opinion, the tests performed to evaluate our approach and described in Section 6.6 confirm that it is really capable of satisfying the lightweightness requirement without sacrificing, if not to a very small extent, result accuracy.

Summarizing, the main contribution of this chapter is an overall procedure capable of extracting interschema properties from structured, semi-structured and unstructured sources. Our procedure is lightweight because it has been specifically conceived to operate on big data. This feature is deeply investigated in the paper, where we analyze its computational requirements and compare them with the one of similar approaches conceived to work on smaller (only) structured and semi-structured data sources. In spite of its lightweightness, the accuracy of our procedure is very satisfying, as witnessed by the quantitative evaluations presented in the paper. An important component of our approach, which could also be extrapolated to other contexts, is the technique for "structuring" unstructured sources whose distinctive peculiarities have been described above.

The rest of this chapter is organized as follows: in Section 6.2, we examine related literature. In Section 6.3, we introduce a source representation model that we exploit in our tasks. In Section 6.4, we show our approach for the construction of a "structured representation" of unstructured data sources. In Section 6.5, we present our interschema property derivation approach. In Section 6.6, we present some experiments that we performed to test our approach.

6.2 Related Literature

6.2.1 Schema matching for structured and semi-structured sources

Schema matching is one of the most investigated topics in past database research. The first schema matching approaches proposed by researchers were manual and operated only on structured databases. Subsequently, researchers proposed semi-automatic or automatic schema matching approaches capable of handling both structured and semi-structured data sources. With the advent of big data, unstructured sources are becoming more and more frequent and important.

Schema matching approaches were thought to consider several kinds of heterogeneity; the most relevant of them are lexicographic, structural and semantic ones. The first deals with names and terms; the second considers type formats, data representation models and structural relationships among concepts; the third regards the meaning of involved data.

Let us see, now, in more detail, an overview of several approaches to perform schema matching from the beginning to the present day.

In [77], an approach to transform structured documents by leveraging schema graph matching is proposed. In particular, an XML schema to map each structured document is defined; for this purpose, some XSLT scripts are automatically generated. In [288], Cupid, a system for deriving interschema properties among heterogeneous sources, is proposed. Cupid leverages two different matchings, namely the *structure* and the *linguistic* ones. In [59], MOMIS, a system supporting querying and information source integration in a semi-automatic fashion, is presented. MOMIS implements a clustering procedure for the extraction of interschema properties. DIKE and XIKE [349, 124, 348], as well as the approaches described in [89, 135], also belong to this generation. An overview of this generation of schema matching approaches can be found in [373, 62].

More recent approaches, which significantly differ from the classical ones, are based on probabilistic methods, applied to networks of schemas [213]. They allow the definition of network-level integrity constraints for matching, as well as the analysis of query/click logs [143, 325], specifying the class of desired user-based schema matching.

In [26], an XML-based schema matching approach conceived to operate on large-scale schemas is presented. This approach leverages Prufer sequences. It performs a two-step activity; during the former step it parses XML schemas in schema trees; during the latter one, it exploits Label Prufer Sequences (LPS) to capture schema tree semantic information. In [332], SMART, a Schema Matching Analyzer and Reconciliation Tool, designed for the detection and the subsequent reconciliation of matching inconsistencies, is proposed. SMART is semi-automatic because it requires the intervention of an expert for the validation of results. In [302], the authors propose an approach to determine the semantic similarity of terms using the knowledge present in the search history logs from Google. For this purpose, they exploit four techniques that evaluate: *(i)* frequent co-occurrences of terms in search patterns; *(ii)* relationships between search patterns; *(iii)* outlier coincidence on search patterns; *(iv)* forecasting comparisons. In [30], a framework for the management of a data lake through the corresponding metadata is proposed. This framework leverages schema matching techniques to identify similarities between the attributes of different datasets. These techniques consider both schemas (specifically, attribute types and dependencies) and instances (specifically, attribute values) [62]. The framework integrates different schema matching approaches proposed in the last years, like graph matching, usage-based matching, document content similarity detection and document link similarity detection. [306] proposes an instance-based approach to find 1-1 schema matches. It combines the semantics provided by Google and regular expressions. It does not work well in a scenario where sources are very heterogeneous and data are stored in their raw way. Another instance-based approach is presented in [217]. It faces the heterogeneity of the different schemas by leveraging an ad-hoc mapping language.

Most schema matching approaches based on similarities often filter out unnecessary matchings and information [358] in such a way as to operate easier and faster.

As we have seen in this overview, schema matching has been widely investigated in the past for very heterogeneous scenarios, and very different approaches have been adopted to reach disparate goals. In this "mare magnum" of approaches, ours is characterized by the following features: *(i)* it has been specifically conceived to handle also unstructured sources; *(ii)* it has been designed to be scalable and, therefore, it is lightweight; *(iii)* it is automatic; *(iv)* in spite of these two last features, it presents a good accuracy, as we will see in Section 6.6.

6.2.2 Approaches to represent unstructured sources

The representation mechanisms of unstructured sources (basically texts) are mainly based on two strategies, namely analysis of contents and analysis of references [430]. The former infers a representation of a document from the corresponding content,

whereas the latter focuses on relationships among documents. Clearly, our interest is mainly on the former strategy, because its objective is similar to the one of our approach.

The most basic approach to represent texts leverages Bags of Words (BOW) [47, 398]. In this case, machine learning techniques are used to identify the set of words that mostly characterizes a text. Some more sophisticated strategies are based on the extraction of sentences [152]. In this case, a text is mapped onto semantic spaces, such as WordNet or Wikipedia. Another strategy is Explicit Semantic Analysis (ESA) [164], which mixes BOW and document references. In ESA, the relatedness between documents is computed by extracting similarities between the concepts identified within them, thanks to the cross-references expressed therein.

An important model in the BOW context is word2vec [309, 310]. This model is based on neural networks. It constructs a vector space and associates each word of the text into examination with a vector in this space in such a way that words sharing common contexts have close corresponding vectors in the vector space. The word2vec model was extended to the doc2vec one [254], which exploits similarities and contextual information of each word to reduce the dimensionality of the vector space. Other approaches reach the same objective (i.e., dimensionality reduction) by means of Latent Semantic Analysis [234], which exploits matrix decomposition methods.

Word-based methods are currently flanked by concept-based ones. As an example, [391, 390] introduce the idea of Bag of Concepts, in place of Bag of Words. In this approach, concepts are generated by disregarding semantic similarities between words. Semantic similarities have been considered only recently [235].

Another relevant set of approaches use ontologies or, in general, external sources of semantics, to generate conceptual representations of documents by matching document terms with ontology concepts (see, for instance, [66, 221, 450, 25]). The performance of these approaches is strongly related to the quality of the adopted external sources. As a consequence, in these approaches, very specific domains can strongly benefit from the availability of dedicated ontologies.

The approaches examined above generally consider only texts; they do not operate with other forms of unstructured sources, such as videos. Furthermore, they terminate with the derivation of keywords or key concepts representing a source. In fact, none of them tries to go a step over, i.e., to define a certain "structure" for an unstructured source, which is one of the objectives of this chapter.

An attempt to define a "structure" for an unstructured source can be found in [291]. This approach generates a rowset with n attributes, i.e., a tabular representation from unstructured data. A single rowset is a set of tuples and is equivalent to a relation in relational databases; logical associations may exist between rowsets, but

these are not explicitly defined. The schema of a rowset may be defined on read. Transformation functions, possibly based on fuzzy logic, are used to properly read the complex unstructured data and map them on the rowset schema. These functions are also exploited to address the data variety issue, by means of an interface for the dataset extraction, which is unified and valid for all the sources. Different transformation functions can be used to map different unstructured data onto the same schema. The content of a rowset depends on the membership function associated with a fuzzy logic and on the possible constraints regarding it. However, data extraction is only one of the steps defined in [291], which develops a general data processing system based on an Extract, Process, and Store (EPS) paradigm.

From the above description, it appears evident that the approach of [291] shares several features with ours; in particular, the purpose of structuring unstructured data is common to both of them. However, the two approaches also present several differences. Indeed, for the structuring task, the approach of [291] strongly depends on user defined transformation functions and on rowset schemas (which are not automatically inferred from the sources). Now, the definition of both the functions and the schema may be difficult for complex sources. Furthermore, mapping more sources on the same schema requires a manual integration step, which, again, may be a difficult task when the number of involved sources is high. On the other hand, querying obtained data sources is particularly effective with the use of fuzzy techniques and the declarative U-SQL query language characterizing the approach of [291]. On the contrary, in our proposal, to perform the structuring of unstructured sources, we leverage network analysis, as well as lexical and string similarities, for automatically deriving a general and uniform schema of different unstructured sources. In fact, as we will see in the following, unstructured sources are "structured" by first representing them as a network, starting from a set of keywords associated with them; then, this structure is enriched by adding arcs that link nodes having lexical or string similarities even if they belong to different sources. As a consequence, it is possible to state that the approach presented in [291] is more effective and flexible in querying data lake contents, but it requires a more complex design phase, with a heavy human intervention, difficult to sustain in presence of numerous data sources. On the contrary, our approach simplifies the structuring phase, because it does not need a preliminary structure to be used as a model, and it does not require a human intervention. On the other side, its querying capabilites are limited to the summarization of unstructured sources provided by the keywords representing them. Therefore, in a certain sense, our approach and the one of [291] can be considered orthogonal.

6.3 A network-based model for uniformly representing structured, semi-structured and unstructured sources

In this section, we present a network-based model for uniformly representing data sources of different formats. This model will be extensively used in the rest of this chapter. In order to understand the peculiarities of our model, we assume to have a set DS of m data sources of interest possibly characterized by different data formats.

$$DS = \{D_1, D_2, \cdots, D_m\}$$

Each data source D_k has associated a rich set $\mathcal{M}_k$ of metadata. We indicate with $\mathcal{M}_{DS}$ the repository of the metadata of all the data sources of DS:

$$\mathcal{M}_{DS} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$$

Given the source D_k, in order to represent the information content stored in $\mathcal{M}_k$, our model starts from a notation typical of XML, JSON and many other semi-structured data models. According to this notation, Obj_k denotes the set of all the objects stored in $\mathcal{M}_k$. Obj_k consists of the union of three subsets:

$$Obj_k = Att_k \cup Smp_k \cup Cmp_k$$

where:

- Att_k denotes the set of the attributes of $\mathcal{M}_k$;
- Smp_k indicates the set of the simple elements of $\mathcal{M}_k$;
- Cmp_k represents the set of the complex elements of $\mathcal{M}_k$.

Here, the meaning of the terms "attribute", "simple element" and "complex element" is the one typical of semi-structured data models.

$\mathcal{M}_k$ can be also represented as a graph:

$$\mathcal{M}_k = \langle N_k, A_k \rangle$$

N_k is the set of the nodes of $\mathcal{M}_k$. There is a node n_{k_j} in N_k for each object o_{k_j} of Obj_k. According to the structure of Obj_k, N_k consists of the union of three subsets:

$$N_k = N_k^{Att} \cup N_k^{Smp} \cup N_k^{Cmp}$$

where N_k^{Att} (resp., N_k^{Smp}, N_k^{Cmp}) denotes the set of the nodes corresponding to Att_k (resp., Smp_k, Cmp_k). There is a biuniovocal correspondence between a node of N_k and an object of Obj_k. Therefore, in the following, we will use these two terms interchangeably. Each node has associated a name that identifies it in the schema which the corresponding element or attribute belongs to.

Let x be a complex element of $\mathcal{M}_k$. We denote by Obj_x the set of the objects directly contained in x and by N_x^{Obj} the set of the corresponding nodes. Finally, let

x be a simple element of $\mathcal{M}_k$. We indicate by Att_x the set of the attributes directly contained in x and by N_x^{Att} the set of the corresponding nodes.

A_k denotes the set of the arcs of $\mathcal{M}_k$. It consists of three subsets:

$$A_k = A'_k \cup A''_k \cup A'''_k$$

where:

- $A'_k = \{(n_x, n_y, L_{xy}) | n_x \in N_k^{Cmp}, n_y \in N_{n_x}^{Obj}\}$; in other words, there is an arc in A'_k from a complex element of $\mathcal{M}_k$ to each object directly contained in it. L_{xy} represents the label of A'_k.
- $A''_k = \{(n_x, n_y, L_{xy}) | n_x \in N_k^{Smp}, n_y \in N_{n_x}^{Att}\}$; in other words, there is an arc in A''_k from a simple element of $\mathcal{M}_k$ to each attribute directly contained in it. L_{xy} represents the label of A''_k.
- $A'''_k = \{(n_x, n_y, L_{xy}) | n_x \in N_k, n_y \in N_k, D_k$ is unstructured, $\sigma(n_x, n_y)$=`true`$\}$. Here, $\sigma(n_x, n_y)$ is a function that receives two nodes and returns `true` if there exists a similarity between n_x and n_y. For instance, $\sigma(n_x, n_y)$ could return `true` if the concepts represented by n_x and n_y are semantically similar or if the names identifying n_x and n_y in the corresponding schema present a high string similarity. L_{xy} represents the label of A'''_k.

As for the label L_{xy} associated with each arc, in the current version of this model, it is `NULL` for the arcs of A'_k and A''_k. However, we do not exclude that, in the future, enrichments of our model might lead us to use this field for storing some knowledge. Instead, L_{xy} has an important meaning for the arcs of A'''_k. In fact, as will be clear in Section 6.5, it is used to denote the strength of the correlation between n_x and n_y.

From an abstract point of view, there is a "fil rouge" linking the three subsets of A_k, which leads to the concept of homophily in Social Network Analysis. Indeed, A'_k, A''_k and A'''_k are the three possible ways to represent the links between a concept and its "direct homophiles", i.e., the other concepts that can contribute to define (at least partially) its meaning.

6.4 Structuring an unstructured source

Our network-based model for uniformly representing and handling data sources with disparate formats is perfectly fitted for semi-structured sources. Indeed, it is sufficient:

- deriving the metadata of the source (if not yet provided) by applying one of the several techniques and tools defined for this purpose w.r.t. the various kinds of format;
- defining a complex element to represent the source as a whole;

- introducing a complex element, a simple element and an attribute for each complex element, simple element and attribute present in the metaschema of the source;
- defining an arc of A'_k from the source to the root of the document;
- introducing an arc of A'_k or A''_k for each relationship existing between the objects composing the source metadata.

Clearly, our model is sufficiently powerful to represent structured data. Indeed, it is sufficient:

- deriving the E/R schema of the source (if not yet provided) by performing a classical database reverse engineering activity;
- defining a complex element to represent the source as a whole;
- introducing a complex element for each entity of the E/R schema and an attribute for each attribute of the schema;
- defining an arc of A'_k from the complex element corresponding to the source to each complex element associated with an entity of the E/R schema;
- introducing an arc of A''_k from an entity to each of its attributes;
- defining an arc of A'_k for each one-to-many relationship of the E/R schema; this arc is from the entity participating to the relationship with a maximum cardinality equal to 1 to the entity participating with a maximum cardinality equal to N;
- representing a many-to-many relationship without attributes as a pair of one-to-many relationships and, then, modeling them accordingly;
- representing a many-to-many relationship R with attributes that connects two entities E_1 and E_2 as an entity having the same attributes as R and linked to E_1 and E_2 by means of two one-to-many relationships; the new entity and the new relationships are then suitably modelled by applying the rules defined in the previous cases.

The highest modeling difficulty regards unstructured data because it is worth avoiding a flat representation consisting of a simple element for each keyword provided to denote the source content. As a matter of fact, this flat representation would make the reconciliation, and the next integration, of an unstructured source with the other semi-structured and structured sources of DS very difficult. This is a very challenging issue to address. In the following, we propose our approach to "structure" unstructured sources. As pointed out in the Introduction, this is one of the main contributions of this chapter. It is in itself a major issue in the current information systems scenario and, at the same time, plays a key role to provide our interschema property derivation approach with the capability of operating on sources with disparate formats.

Our approach assumes that each unstructured source into consideration (e.g., a video, an audio, an image, a text) is provided with a list of keywords describing it[1]. They will play a key role, as will be clarified in the following. We observe that this assumption is not particularly strong or out of place. As a matter of fact, in the reality, most video, image or audio providers associate a list of keywords (sometimes, in the form of tags) with the contents they deliver. As for text, representing keywords can be also easily derived through suitable techniques, like TF-IDF [299].

Our approach consists of four phases, namely: *(1)* creation of nodes; *(2)* management of lexical similarities; *(3)* management of string similarities; *(4)* management of (temporary) duplicated arcs. We describe these phases below.

- **Phase 1: Creation of nodes.** During this phase, our approach creates a complex node representing the source as a whole and a simple node for each keyword[2]. Furthermore, it adds an arc of A'_k from the node associated with the source to any node corresponding to a keyword. Initially, there is no arc between two keywords. To determine the arcs to add, the next phases are necessary.
- **Phase 2: Management of lexical similarities.** During this phase, our approach handles lexical similarities. For this purpose, it leverages a suitable thesaurus. Taking the current trends into account, this thesaurus should be a multimedia one; for this purpose, in our experiments, we have adopted BabelNet [326]. In particular, our approach adds an arc of A'''_k from the node n_{k_1}, corresponding to the keyword k_1, to the node n_{k_2}, corresponding to the keyword k_2, and vice versa, if k_1 and k_2 have at least one common lemma[3] in the thesaurus. Furthermore, it transforms the nodes n_{k_1} and n_{k_2} from simple to complex. The new arcs have a label corresponding to the number of common lemmas for k_1 and k_2 in the thesaurus.
- **Phase 3: Management of string similarities.** During this phase, our approach derives string similarities and states that there exists a similarity between two keywords k_1 and k_2 if the string similarity degree $kd(k_1, k_2)$, computed by applying a suitable string similarity metric on k_1 and k_2, is "sufficiently high" (see below). In this case, it adds an arc of A'''_k from n_{k_1} to n_{k_2}, and vice versa. Both the two arcs

[1] Here, we assume that the list is ordered and the order is the one in which the keywords appear in the list.

[2] Here and in the following, to make the presentation smoother, we use the term "complex node" to indicate a node belonging to N_k^{Cmp} and the term "simple node" to denote a node of N_k^{Smp}. Furthermore, we use the term "source" (resp., "keyword") to denote both the source (resp., a keyword) and the corresponding node associated with it.

[3] In this chapter, we use the term "lemma" according to the meaning it has in BabelNet [326]. Here, given a term, its lemmas are other objects (terms, emoticons, etc.) that contribute to specify its meaning.

have $kd(k_1, k_2)$ as their label. We have chosen N-Grams [241] as string similarity metric because we have experimentally seen that it provides the best results in our context. In particular, we have selected bi-grams as the best trade-off between accuracy and costs. In fact, mono-grams would require a lower cost but they would also return a lower accuracy than bi-grams. By contrast, tri-grams would guarantee a very high accuracy but at the expense of the computational cost, which would be excessive. Again, if n_{k_1} and n_{k_2} are simple nodes, our approach transforms them into complex ones.

Now, we illustrate in detail what "sufficiently high" means and how our approach operates. Let $KeySim$ be the set of the string similarities for each pair of keywords of the source into consideration. Each record in $KeySim$ has the form $\langle k_i, k_j, kd(k_i, k_j)\rangle$. Our approach first computes the maximum keyword similarity degree kd_{max} present in $KeySim$. Then, it examines each keyword similarity registered therein. Let $\langle k_1, k_2, kd(k_1, k_2)\rangle$ be one of these similarities. If $((kd(k_1, k_2) \geq th_k \cdot kd_{max})$ and $(kd(k_1, k_2) \geq th_{kmin}))$, which implies that the keyword similarity degree between k_1 and k_2 is among the highest ones in $KeySim$ and that, in any case, it is higher than or equal to a minimum threshold, then it concludes that there exists a similarity between n_{k_1} and n_{k_2}. We have experimentally set $th_k = 0.70$ and $th_{kmin} = 0.50$.

Observe that the choice to consider string similarities, in particular the one to adopt N-Grams as the technique for detecting string similarities, makes our approach robust against mispelling errors possibly present in the keywords. In fact, as shown in [194], N-Grams is well suited to handle also this kind of error.

- **Phase 4: Management of (temporary) duplicated arcs.** This phase is devoted to handle the possible simultaneous presence of both lexical and string similarities for the same pair of keywords. Indeed, it may occur that, for a pair of nodes n_{k_1} and n_{k_2}, there are two arcs from n_{k_1} to n_{k_2} belonging to A_k''' and generated by both lexical and string similarities, and two arcs from n_{k_2} to n_{k_1}. In this case, the two arcs from n_{k_1} to n_{k_2} corresponding to these two forms of similarities, must be merged in only one arc, which has associated a label denoting both the number of common lemmas between k_1 and k_2 in BabelNet and the value of $kd(k_1, k_2)$. The same happens for the two arcs from n_{k_2} to n_{k_1}.

From this description, it emerges that, at the end of the four phases, given two nodes n_{k_1} and n_{k_2}, four cases may exist, namely:

1. There is no arc from n_{k_1} to n_{k_2}.
2. A pair of arcs derived from a lexical similarity links them. In this case, the two arcs actually coincide (also in their labels); therefore, one of them can be removed. Note

that the choice of the arc to be removed has deep implications in the definition of the topology of the corresponding network. Indeed, one of the two nodes involved (i.e., the source node of the maintained arc) will be certainly a complex node, whereas the other one may be a simple node (if no other arc starts from it) or a complex node (if at least another arc, different from the removed one, starts from it). In turn, the topology of the network has implications in the nature and the quality of the interschema properties that can be extracted, as will be clear in Section 6.5. Therefore, it is appropriate that the choice of the arc to be removed is not random and that a clear rule guiding it is defined. The rule that we chose for our approach is the following: given a pair of arcs between two nodes n_{k_1}, corresponding to the keyword k_1, and n_{k_2}, corresponding to the keyword k_2, with k_1 preceding k_2 in the list of keywords associated with the source D_k, the arc from n_{k_1} to n_{k_2} is maintained and the one from n_{k_2} to n_{k_1} is removed.

3. A pair of arcs derived from a string similarity links them. As in the previous case, the two arcs coincide and one of them is removed. The policy adopted to determine the arc to remove is the same as the one followed in the previous case.
4. A pair of arcs derived from Phase 4 links them. As in the previous case, the two arcs coincide and one of them is removed.

Actually, arc labels introduced above are not necessary in our approach for the extraction of semantic relationships described in Section 6.5. However, we have decided to maintain them in our model because we aim at providing an approach to "structure" unstructured sources that is general and that may be adopted in several future applications, some of which could benefit from this information.

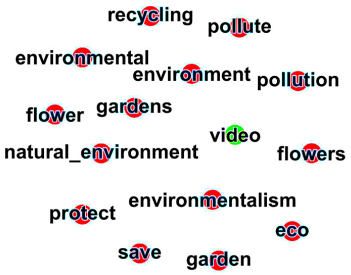
Moreover, we point out that, in the prototype implementing our approach, in order to increase its efficiency, we directly added only one arc, namely (n_{k_1}, n_{k_2}), during Phases 2, 3 and 4, instead of adding two arcs and of removing one of them at the end of the four phases.

6.4.1 Example

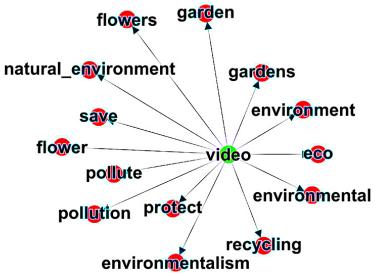
In this section, we propose an example of how our approach to construct a "structured" representation of an unstructured source operates. In particular, the unstructured source into consideration is a video, which talks about environment and pollution. As we said before, for each unstructured source, our approach begins from a list of keywords representing its content. In order to keep our description simple and clear, in this example, we assume that our video has a limited number of keywords, namely the ones shown in Figure 6.1.

Our approach starts with Phase 1. As we can see in Figure 6.1(a), during this phase, it constructs a graph having a node for each keyword. A further node is added

(a)



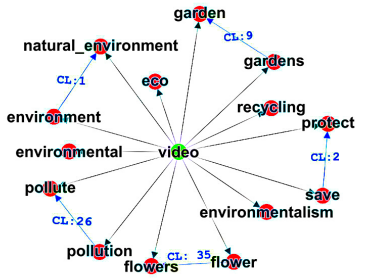
(b)



(c)

	Lemmas
Save	blown_save, bring_through, carry_through, conditional_salvation, deliver, economise, economize, file, hold_open, individual_salvation, kaligtasan, **keep**, keep_open, lay_aside, make_unnecessary, noar, noarus, **preserve**, proprietary_salvation, pull_through, redeem, relieve,rescued, salvage, salvation, sava, save, save_situation, save_up, saved, saving, savior, saviour, savus, sin_and_salvation, spare, special_salvation, svop, szva, tough_save, write
Protect	defend, guard, **keep**, **preserve**, protect, protect_(organisation), protect_(organization), protect_(political_organization), The_national_association_to_protect_children

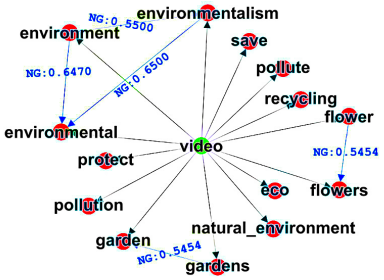
(d)



(e)

k_1	k_2	$kd(k_1, k_2)$
environmentalism	environmental	0.6500
environmental	environment	0.6470
environment	environmentalism	0.5500
flower	flowers	0.5454
gardens	garden	0.5454

(f)



(g)

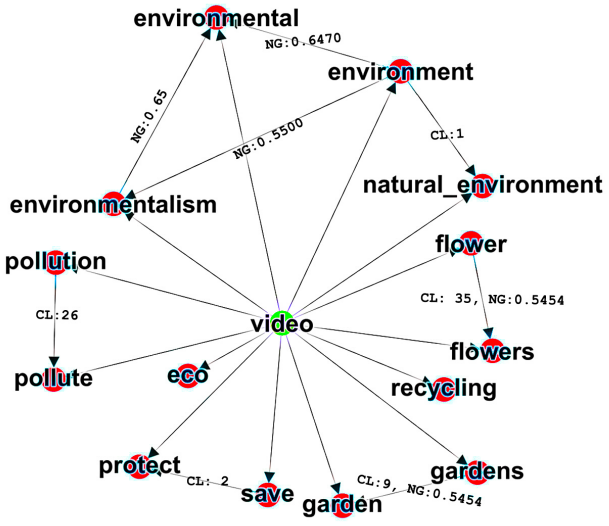


Fig. 6.1. Graphical representation of our approach to derive a "structure" for an unstructured source

to represent the video as a whole; nodes representing keywords are colored in red, whereas the other one is colored in green. Following our strategy, in Figure 6.1(b), we added an arc from the node representing the whole video to each node associated with a keyword.

Now, Phase 2 starts. During this phase, our approach uses a thesaurus. In our example, we leveraged BabelNet. In particular, let k_1 and k_2 be two keywords of Figure 6.1(a) having at least one common lemma in BabelNet. An arc is added from the node n_{k_1}, associated with k_1, to the node n_{k_2}, associated with k_2, and vice versa. In Figure 6.1(c), we show two keywords ("Save" and "Protect") and the corresponding lemmas in BabelNet. Common lemmas (i.e., "keep" and "preserve") are in bold. Since "Save" and "Protect" have at least one common lemma, an arc is added between the corresponding nodes in Figure 6.1(d)[4]. This arc is highlighted in blue. Each arc has a label representing the number of common lemmas between the corresponding keywords in BabelNet.

After having examined lexical similarities, Phase 2 terminates and our approach proceeds with Phase 3, which leverages string similarities. In particular, let k_1 and k_2 be two keywords of Figure 6.1(a) having a string similarity degree higher than or equal to $th_k \cdot kd_{max}$ and, at the same time, higher than or equal to th_{kmin}. An arc is added from the node n_{k_1}, corresponding to k_1, to the node n_{k_2}, corresponding to k_2. In Figure 6.1(e), we report the pairs of keywords that satisfy this feature. In Figure 6.1(f), we added an arc for each pair of keywords of Figure 6.1(e). Here, to better highlight them, we have omitted the arcs constructed during Phase 2. Again, these arcs are highlighted in blue. Each arc has a label representing the string similarity degree (computed by means of N-Grams) between the corresponding keywords.

Finally, in Figure 6.1(g), Phase 4 of our approach combines the arcs derived in Phases 2 and 3. In particular, it may happen that, for a pair of keywords (see, for instance, the keywords "garden" and "gardens"), two arcs have been generated, one in Figure 6.1(d) and one in Figure 6.1(f). In this case, in Figure 6.1(g), the two arcs are substituted by only one arc, representing both of them. The label of this arc reports the label of both the original ones.

6.5 Extracting interschema properties from disparate sources

We are now ready to illustrate our strategy for uniformly extracting interschema properties from structured, semi-structured and unstructured sources. Here, we assume that the content of the sources of interest is represented by means of the model

[4] Here, we have directly added only one arc between "Save" and "Protect", instead of adding two arcs and removing one of them later, after the four phases.

described in Section 6.3, and that our approach to "structure" unstructured sources, described in Section 6.4, has been already applied on all unstructured sources.

Before delving into a detailed description of our approach, a discussion about the role played by source metadata, and about the consequences of this role, is in order. Indeed, as previously pointed out, our approach assumes that some metadata are available for each structured, semi-structured and unstructured source. This assumption is important because both our approach for structuring unstructured sources and our approach for extracting interschema properties use these metadata. It is, then, of outmost importance to analyze the possible issues (and the corresponding solutions) in obtaining good quality metadata, when they are not directly provided with the sources, and the impact that they have on the results returned by our approach.

Metadata generation received much attention in the literature. According to [23], metadata relative to a data source are currently generated by crawlers, by professional metadata creators, or, finally, by source creators. Generating metadata by means of automatic crawlers has great advantages, such as low cost and high efficiency; however, in some cases, the quality of generated metadata could be poor. In this context, it could be extremely useful the support of several mechanisms for controlling the quality of metadata, as well as the aid of metadata professionals, such as cataloguers and indexers; these are people who have had a formal training and are efficient in using metadata. Generally, they produce high-quality metadata. However, it has been observed that, in some cases, even metadata generated by professionals or by source authors may have poor quality and might hamper, rather than aid, the usage of the corresponding sources. This happens because most authors have little previous knowledge on metadata creation [23].

As pointed out in [353], the widespread adoption of several mechanisms for controlling the quality of metadata witnesses a strong awareness of the importance of having high-quality metadata at disposal. However, despite the relevance and the impact of metadata quality are universally recognized in the literature, there is no agreement yet on what metadata quality actually means. This implies, among the other things, the impossibility of defining systematic approaches to its automatic measurement and enhancement [432]. Metadata quality assurance should be verified simultaneously to metadata creation [352]. Indeed, a poor quality of metadata negatively affects the performance of systems using them and the overall user satisfaction. Quality assurance procedures are generally complemented by manual quality review and, if necessary, by the assistance of the technical staff during the process of metadata creation. Other mechanisms, such as metadata creation guidelines (sometimes embedded into the metadata creation system) and metadata generation tools, are on the rise.

The great relevance given to the metadata quality improvement is observed in the study presented in [226]. Here, the authors introduce a quality measure and analyze the metadata quality in the Europeana context over the years. They observe that the metadata quality improves not only in new collections but also in the same collection over the years.

As pointed out in [353], in the metadata generation process, accuracy and consistency are prioritized over completeness, whereas the semantics of metadata elements is perceived to be less important. In principle, this might be an issue for our approach, since it strongly relies on semantics. The authors of [353] also point out that semantic overlaps and ambiguities are by far the two most critical factors. Actually, as our approach exploits thesauruses, string, and semantic similarities to relate keywords, these negative factors are significantly mitigated.

After this important discussion about the metadata of the involved sources, we can start our discussion about the derivation of interschema properties. We recall that, in the current big data scenario, any interschema property extraction strategy must be lightweight. For this reason, in our effort to define a new approach for this task, we avoided highly complex choices, such as the fixpoint computation characterizing DIKE [349, 348] and XIKE [124], or the clustering-based computation characterizing MOMIS [60], or, again, the wide range of parameter computation characterizing Cupid [288]. These choices, as well as most of the other ones present in the past approaches proposed for reconciling and integrating structured and semi-structured data sources (e.g., the construction of a data warehouse) [373, 62], would certainly return very accurate results. However, their speed is incompatible with the one required in many current applications, which must allow the derivation of semantic relationships "on-the-fly" from a very high number of data sources, most of which are unstructured, i.e., in a format not considered by classic approaches. As a consequence, our strategy must necessarily privilege quickness over accuracy even if, clearly, accuracy must be high. In Section 6.6, we will see if, and how, this issue has been addressed.

Our strategy consists of two phases; the former computes the semantic similarity degree of each pair of objects stored in the metadata of the involved sources. The latter derives semantic relationships between the same objects starting from the results returned by the former.

6.5.1 Semantic similarity degree computation

Our approach to semantic similarity degree computation consists of three steps, namely:

- basic similarity computation;

- standard similarity computation;
- refined similarity computation.

In the next subsections, we illustrate these three steps in detail.

Basic similarity computation

Basic similarities consider only lexicon (determined with the support of suitable thesauruses, such as BabelNet [326] and WordNet [311], and string similarity metrics, such as N-Grams [241]), and object types.

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The basic similarity degree $bs(x_1, x_2)$ between x_1 and x_2 can be computed as:

$$bs(x_1, x_2) = \omega \cdot \sigma_L(x_1, x_2) + (1 - \omega) \cdot \sigma_T(x_1, x_2)$$

In other words, the basic similarity degree between x_1 and x_2 can be computed as a weighted mean of two components. The former, σ_L, returns their lexical similarity, whereas the latter, σ_T, specifies the similarity of their types. ω is a weight belonging to the real interval $[0, 1]$ and used to tune the importance of σ_L w.r.t. σ_T. We have experimentally set ω to 0.90.

σ_L can be directly detected from a thesaurus. In our experiments, we used WordNet in the first beat, because it provides the similarity degree between the two objects, and BabelNet, when WordNet did not provide any result. Since this last thesaurus does not return the similarity degree of two objects that it considers similar, we coupled BabelNet with a suitable string similarity metric (in particular, N-Grams). This last is applied to the objects and the corresponding lemmas returned by BabelNet; obtained results are, then, combined to compute the lacking similarity degree. Furthermore, in very specific application contexts, specialized thesauruses could be used.

σ_T is defined as follows:

$$\sigma_T = \begin{cases} 1 & \textit{if } (x_1 \in Cmp_1 \textit{ and } x_2 \in Cmp_2) \textit{ or } (x_1 \in Smp_1 \textit{ and } x_2 \in Smp_2) \textit{ or} \\ & \quad (x_1 \in Att_1 \textit{ and } x_2 \in Att_2) \\ 0.5 & \textit{if } (x_1 \in Cmp_1 \textit{ and } x_2 \in Smp_2) \textit{ or } (x_1 \in Smp_1 \textit{ and } x_2 \in Cmp_2) \textit{ or} \\ & \quad (x_1 \in Smp_1 \textit{ and } x_2 \in Att_2) \textit{ or } (x_1 \in Att_1 \textit{ and } x_2 \in Smp_2) \\ 0 & \textit{otherwise} \end{cases}$$

Standard similarity computation

Standard similarities take both basic similarities and the neighbors of the involved objects into account.

Let D_k be a source of the set DS of the sources of interest, let $\mathcal{M}_k = \langle N_k, A_k \rangle$ be the corresponding set of metadata, let Obj_k be the set of the objects of $\mathcal{M}_k$. The set $nbh(x)$ of the neighbors of an object $x \in Obj_k$ is defined as:

$$nbh(x) = \{y | y \in Obj_k, (n_x, n_y) \in A_k\}$$

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The standard similarity degree $ss(x_1, x_2)$ between x_1 and x_2 can be computed as follows:

- If both $nbh(x_1) = \emptyset$ and $nbh(x_2) = \emptyset$, then $ss(x_1, x_2) = bs(x_1, x_2)$ [5].
- If either $nbh(x_1) = \emptyset$ and $nbh(x_2) \neq \emptyset$ or $nbh(x_2) = \emptyset$ and $nbh(x_1) \neq \emptyset$, then $ss(x_1, x_2) = f_p \cdot bs(x_1, x_2)$. Here, f_p is a factor, whose possible values belong to the real interval $[0, 1]$, which "penalizes" the value obtained for basic similarities. Indeed, these are the only similarities that we can compute and, therefore, we must base our standard similarity computation on them. However, we must consider that the sets of neighbors of x_1 and x_2 have different features, because one of them is empty and the other one is not empty, and this fact must be taken into account. We have experimentally set $f_p = 0.85$.
- In all the other cases, i.e., if $x_1 \in (Smp_1 \cup Cmp_1)$ and $x_2 \in (Smp_2 \cup Cmp_2)$, then $ss(x_1, x_2)$ can be computed as follows:
  1. $nbh(x_1)$ and $nbh(x_2)$ are determined.
  2. A bipartite graph, whose nodes are the ones of $nbh(x_1)$ and $nbh(x_2)$, is constructed.
  3. For each pair (p, q), such that $p \in nbh(x_1)$ and $q \in nbh(x_2)$, an arc is added in the bipartite graph; the weight of this arc is set to $bs(p, q)$.
  4. The maximum weight matching is computed on this bipartite graph. Let A_M be the set of the returned arcs. Then:

$$ss(x_1, x_2) = \begin{cases} \frac{2 \cdot \sum_{(p,q) \in A_M} bs(p,q)}{|nbh(x_1)| + |nbh(x_2)|} & \textit{if neither } D_1 \textit{ nor } D_2 \textit{ are unstructured} \\ \\ \frac{2 \cdot \sum_{(p,q) \in A_M} bs(p,q)}{2 \cdot min(|nbh(x_1)|, |nbh(x_2)|)} & \textit{otherwise} \end{cases}$$

  In this formula, if neither D_1 nor D_2 are unstructured, $ss(x_1, x_2)$ returns the value of an objective function that takes into account how many nodes of $nbh(x_1)$ and $nbh(x_2)$ are linked by basic similarity relationships and how strong these relationships are. Furthermore, the objective function penalizes the presence of dangling nodes, i.e., nodes of $nbh(x_1)$ or $nbh(x_2)$ that do not participate to the maximum weight matching.

[5] For instance, this happens when both x_1 and x_2 are attributes; indeed, the nodes corresponding to attributes do not have outgoing arcs.

If D_1 and/or D_2 are unstructured, then it is necessary to consider that, even if our approach performed a "structuring" task, its final structure is limited, if compared with the rich structure characterizing the other kinds of source. As a consequence, the sets of neighbors of the nodes belonging to unstructured sources are generally much smaller than the ones characterizing the other kinds of source. Therefore, in this case, using the same objective function adopted when neither D_1 nor D_2 are unstructured would not take this important feature into account, and the overall result would be biased. To address this issue, if D_1 and/or D_2 are unstructured, in the denominator of $ss(x_1, x_2)$ we consider the minimum size between $|nbh(x_1)|$ and $|nbh(x_2)|$, clearly multiplied by 2 to indicate the maximum number of nodes that could be linked by a similarity relationship in this situation.

Refined similarity computation

Refined similarities are based on standard similarities (for simple and complex objects), basic similarities (for attributes) and object neighbors.

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The refined similarity degree $rs(x_1, x_2)$ between x_1 and x_2 can be computed as follows:

- If $nbh(x_1) = \emptyset$ and/or $nbh(x_2) = \emptyset$, then $rs(x_1, x_2) = ss(x_1, x_2)$.
- Otherwise, if $x_1 \in (Smp_1 \cup Cmp_1)$ and $x_2 \in (Smp_2 \cup Cmp_2)$, then $rs(x_1, x_2)$ is obtained by applying the same four steps described for $ss(x_1, x_2)$ with the only difference that, in Step 3, the weight of the arc (p, q), such that $p \in nbh(x_1)$ and $q \in nbh(x_2)$, is set to $ss(p, q)$, and no more to $bs(p, q)$. In other words, while standard similarity computation leverages basic similarities, refined similarity computation is based on standard similarities.

Clearly, from a theoretical point of view, it would be possible to perform other refinement steps. In this case, at the i^{th} refinement step, the similarities would be computed starting from the ones obtained at the $(i-1)^{th}$ step, by setting these last ones as the weights of the arcs of the bipartite graph. However, the advantages in accuracy that these further refinement steps could produce do not justify the computational costs introduced by them (see Section 6.6), especially in an agile and lightweight context, such as the one characterizing the big data scenario.

6.5.2 Semantic relationship detection

The derivation of semantic relationships among the objects of the sources of DS represents the second phase of our strategy. It takes the refined semantic similarities among the objects of DS as input. The semantic relationships that it can return are the following:

- *Synonymies*: A synonymy between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ exists if they have a high similarity degree, the same type (i.e., both of them are complex objects or simple objects or attributes) and (possibly) different names.
- *Type Conflicts*: A type conflict between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ exists if they have a high similarity degree but different types.
- *Overlappings*: An overlapping exists between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ if they have (possibly) different names, the same type and an intermediate similarity degree, in such a way that they can be considered neither synonymous nor distinct.
- *Homonymies*: A homonymy between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ exists if they have the same name and the same type but a low similarity degree.

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Finally, let $RefSim_{12}$ be the set of refined similarities involving the objects of Obj_1 and Obj_2.

First, our approach computes the maximum refined similarity degree rs_{max} present in $RefSim_{12}$. Then, it examines each similarity $\langle x_1, x_2, rs(x_1, x_2)\rangle$ registered in $RefSim_{12}$ and verifies if a semantic relationship exists between the corresponding objects as follows:

- If $(rs(x_1, x_2) \geq th_{Syn} \cdot rs_{max})$ and $(rs(x_1, x_2) \geq th_{min})$, which implies that the refined similarity degree between x_1 and x_2 is among the highest ones in $RefSim_{12}$ and, in any case, higher than or equal to a minimum threshold, then:
  - if x_1 and x_2 have the same type, it is possible to conclude that a synonymy exists between them;
  - if x_1 and x_2 have different types, it is possible to conclude that a type conflict exists between them.
- If $(rs(x_1, x_2) < th_{Syn} \cdot rs_{max})$ and $(rs(x_1, x_2) \geq th_{Ov} \cdot rs_{max})$ and $(rs(x_1, x_2) \geq th_{min})$, which implies that the refined similarity degree between x_1 and x_2 is higher than or equal to a minimum threshold, it is not among the highest ones in $RefSim_{12}$, but it is significant, then:

  - if x_1 and x_2 have the same type, it is possible to conclude that an overlapping exists between them.

- If $(rs(x_1, x_2) < th_{Hom} \cdot rs_{max})$ and $(rs(x_1, x_2) < th_{max})$, which implies that the refined similarity degree between x_1 and x_2 is among the lowest ones in $RefSim_{12}$ and, in any case, lower than a maximum threshold, then:
  - if x_1 and x_2 have the same name and the same type, it is possible to conclude that a homonymy exists between them.

Here, th_{Syn}, th_{min}, th_{Ov}, th_{Hom} and th_{max} have been experimentally set to 0.85, 0.50, 0.65, 0.25 and 0.15, respectively.

As pointed out in the Introduction, the knowledge of interschema properties is very relevant for several applications, for instance source integration, source querying, data warehouse and/or data lake construction, data analytics, and so forth. As an example, as far as source integration is concerned:

- If a synonymy exists between $x_1 \in Obj_1$ and $x_2 \in Obj_2$, then x_1 and x_2 must be merged in a unique object, when the integrated schema is constructed.
- If a homonymy exists between x_1 and x_2, then it is necessary to change the name of x_1 and/or x_2, when the integrated schema is constructed.
- If an overlapping exists between x_1 an x_2, then it is necessary to restructure the corresponding portion of network. Specifically, a node x_{12}, representing the "common part" of x_1 and x_2, is added to the network. Furthermore, each pair of arcs (x_1, x_T) and (x_2, x_T), starting from x_1 and x_2 and having the same target x_T, is substituted by a unique arc (x_{12}, x_T). Finally, an arc from x_1 to x_{12} and another arc from x_2 to x_{12} are added to the network.
- If a type conflict exists between x_1 and x_2, then it is necessary to change the type of x_1 and/or x_2 in such a way as to transform the type conflict into a synonymy. Then, it is necessary to handle this last relationship by applying the corresponding integration rule seen above.

The way of proceeding described above can be extended to the detection of hyponymies. In particular, the extension already proposed in [346] for structured and semi-structured data can be probably adapted to this scenario. We plan to investigate this issue in the future. Finally, an analogous way of proceeding can be performed when querying or other activities must be carried out on a set of sources of interest.

An example case

In this section, we provide an example of the behavior of our approach to the extraction of semantic relationships. To fully illustrate its potentialities, we derive these rela-

tionships between objects belonging to an unstructured source and a semi-structured one.

The unstructured source is a video. The corresponding keywords are reported in Table 6.1. Its "structured" representation, in our network-based model, obtained after the application of the approach described in Section 6.4, is reported in Figure 6.2. The semi-structured source is a JSON file whose structure is shown in Figure 6.3. Its representation in our network-based model is reported in Figure 6.4.

Keywords
$video, reuse, flower, easy, tips, plastic, simple, environment, pollution, garbage, wave,$ $recycle, reduce, pollute, help,$ $natural\_environment, educational, green, environment\_awareness, bike, life,$ $environmentalism, planet,$ $earth, climate, clime, save, nature, environmental, gardens, power, recycling,$ $garden, protect, flowers,$ $eco, fine\_particle, o_3, atmospheric\_condition, ocean, metropolis, weather,$ $spot, waving, aurora$

Table 6.1. Keywords of the unstructured source of our interest

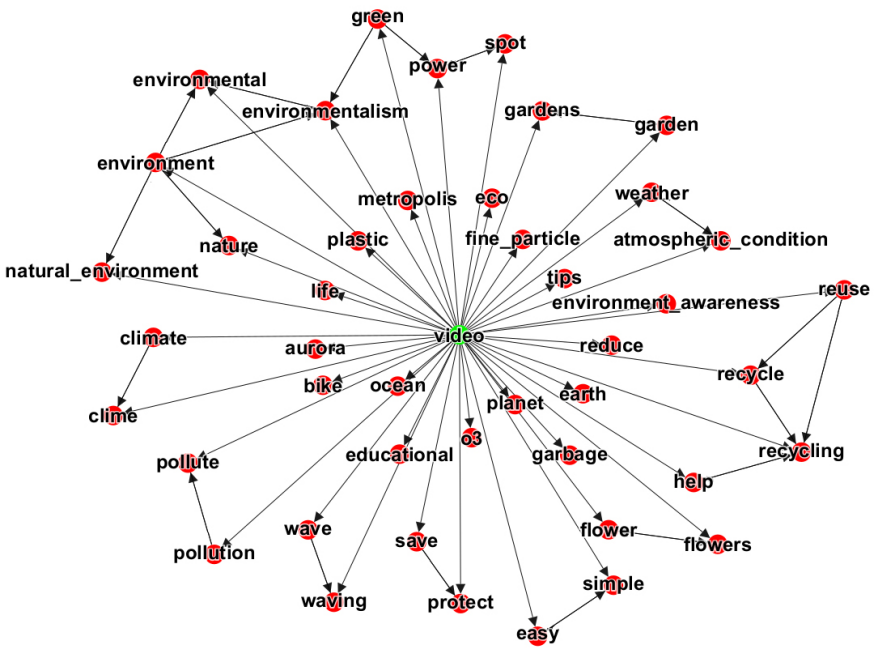


Fig. 6.2. Representation, in our network-based model, of the unstructured source of our interest

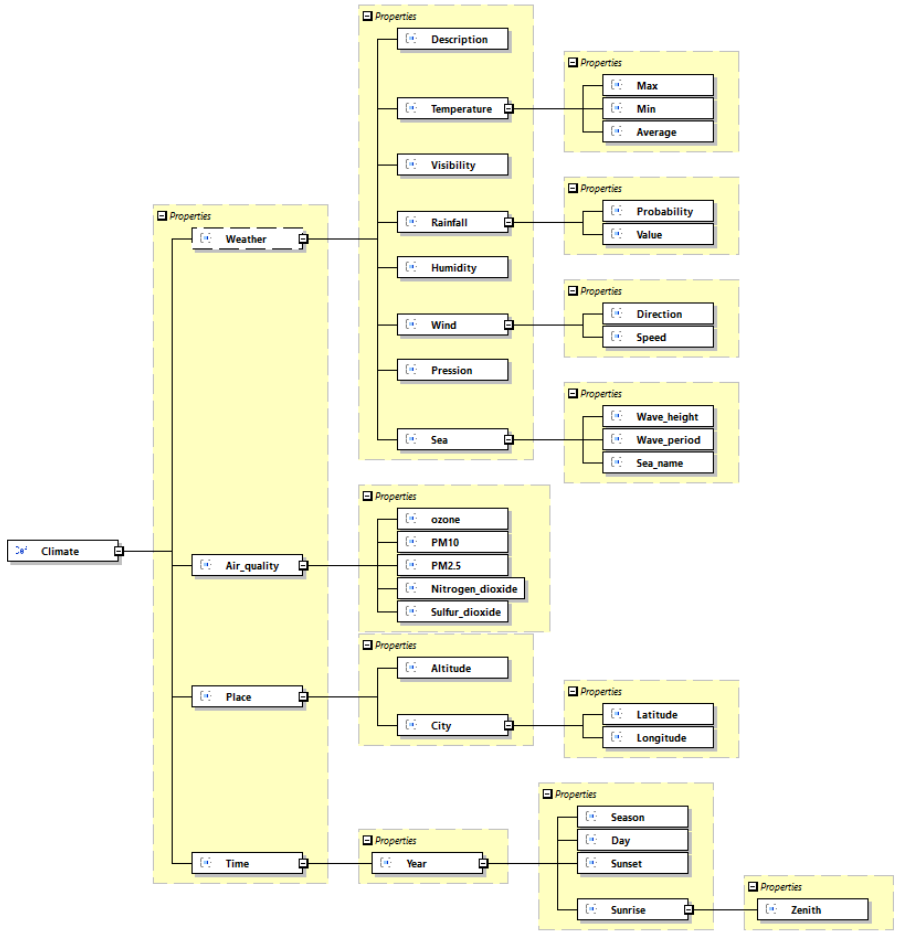


Fig. 6.3. Structure of the JSON file associated with the semi-structured source of our interest

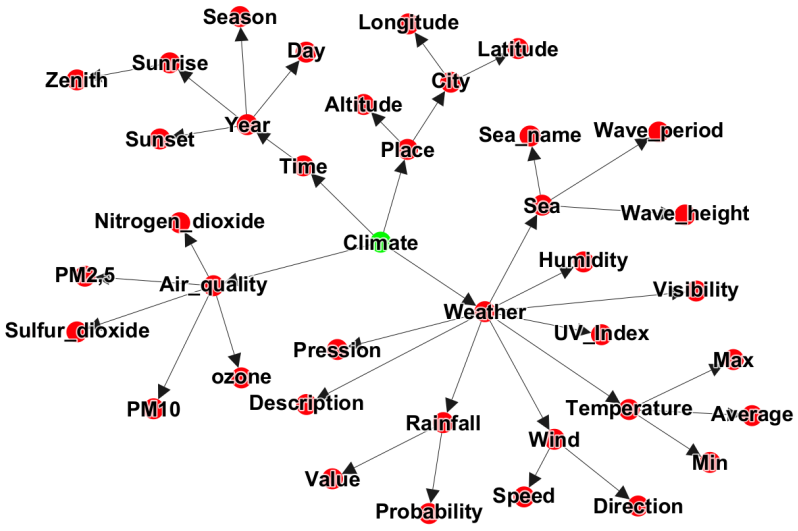


Fig. 6.4. Representation, in our network-based model, of the semi-structured source of our interest

By applying the first phase of our approach we obtained the refined semantic similarity degrees between all the possible pairs of nodes (n_U, n_S), such that n_U

belongs to the unstructured source and n_S belongs to the semi-structured one. To give an idea of these similarity degrees, in Figure 6.5, we report their distribution in a semi-logarithmic scale. From the analysis of this figure, we can observe that a very few number of pairs have a significant similarity degree, which could make them eligible to be selected for synonymies, type conflicts and overlappings. At a first glance, this trend appeared correct and intuitive, even if this conclusion had to be confirmed or rejected by a much deeper analysis (see below).

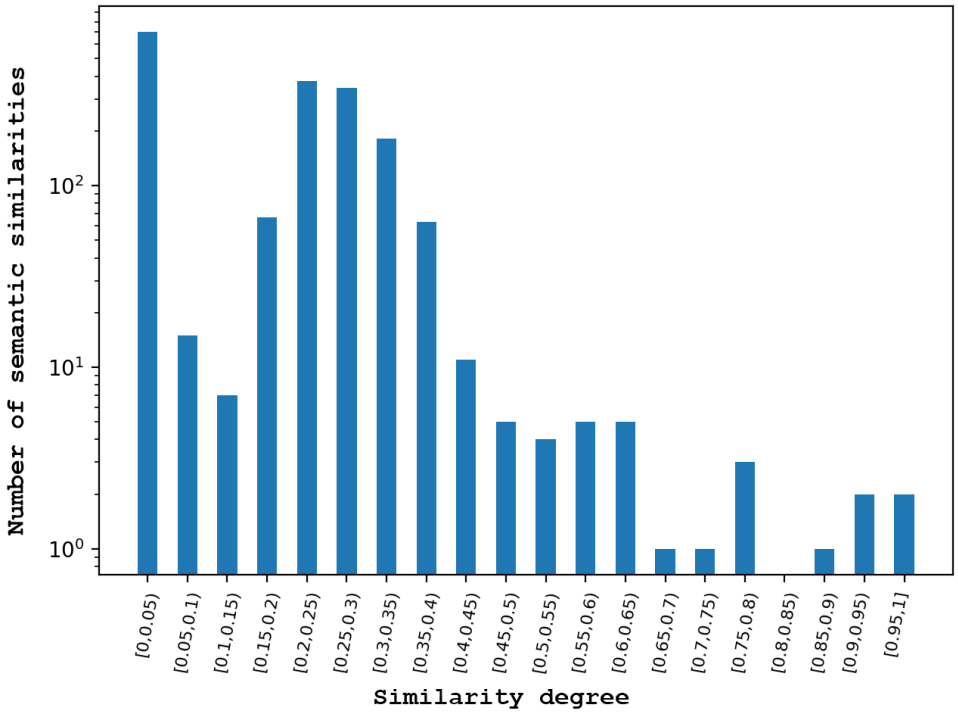


Fig. 6.5. Distribution, in a semi-logarithmic scale, of the values of the the semantic similarity degrees of the objects belonging to the two sources of interest

By applying the second phase of our approach, we obtained the synonymies, the type conflicts and the overlappings reported in Tables 6.2 - 6.4. Instead, as for this pair of sources, we found no homonymies.

Semi-Structured Source Node	*Unstructured Source Node*
climate	*climate*
climate	*clime*

Table 6.2. Derived synonymies between objects of the two sources of interest

We asked a human expert to validate these results. At the end of this task, he reported the following considerations:

Semi-Structured Source Node	*Unstructured Source Node*
$pm10$	$fine\_particle$
$ozone$	o_3

Table 6.3. Derived type conflicts between objects of the two sources of interest

Semi-Structured Source Node	*Unstructured Source Node*
sea	*ocean*
city	*metropolis*
sunrise	*aurora*
place	*spot*
wind	*tips*
$sulfur\_dioxide$	*garbage*
weather	*clime*

Table 6.4. Derived overlappings between objects of the two sources of interest

- The synonymies provided by our approach are correct. No further synonymy can be manually found in the two considered sources.
- The type conflicts provided by our approach are correct. No further type conflict can be manually found in the two sources.
- The overlappings provided by our approach are correct, except for the one linking "wind" and "tips", which actually represents two different concepts. A very interesting overlapping found by our approach is the one between "sulfur_dioxide" and "garbage", in that, even if they represent two seemingly different concepts, both of them denote harmful substances. Some further overlappings could be manually found in the two sources into consideration (for instance, the one between "climate" and "environment"), even if they are semantically weak, and considering them as overlappings or as distinct concepts is subjective.

6.6 Experiments

Our test campaign had four main purposes, namely: *(i)* evaluating the performance of our interschema property derivation approach when applied to the scenario for which it was thought, *(ii)* evaluating the pros and the cons of this approach w.r.t. analogous ones thought for structured and semi-structured sources, *(iii)* evaluating its scalability, and *(iv)* evaluating the role of our approach for structuring unstructured sources. We describe these four experiments in the next subsections.

6.6.1 Overall performances of our approach

To perform our experiments, we constructed a set DS of data sources consisting of 2 structured sources, 4 semi-structured ones (2 of which were XML sources and 2 were JSON ones), and 4 unstructured ones (2 of which were books and 2 were videos). All these sources stored data about environment and pollution. To describe unstructured sources, we considered a list of keywords for each of them. These keywords were derived from Google Books, for books, and from YouTube, for videos. The interested reader can find the schemas, in case of structured and semi-structured sources, and the keywords, in case of unstructured sources, at the address `http://daisy.dii.univpm.it/dl/datasets/dl1`. The password to type is "`za.12&;lq74:#`".

It could appear that taking only 10 sources is excessively limited. However, we made this choice because we wanted to fully analyze the behavior and the performance of our approach and, as it will be clear below, this requires the human intervention for verifying obtained results. This intervention would have become much more difficult with a higher number of sources to examine. At the same time, our test set is fully scalable. As a consequence, an interested reader, starting from the data sources provided at the address `http://daisy.dii.univpm.it/dl/datasets/dl1`, can construct a data set with a much higher number of sources, if necessary.

For our experiments, we used a server equipped with an Intel I7 Dual Core 5500U processor and 16 GB of RAM with the Ubuntu 16.04.3 operating system. Clearly, the capabilities of this server were limited. However, they were adequate for the (small) data set DS we have chosen to use in our tests.

As the first task of our experiment, we represented the metadata of all the sources by means of the data model described in Section 6.3. Then, we applied the approach described in Section 6.4 to (at least partially) "structure" the unstructured sources of our test data set. Finally, we extracted semantic relationships existing between all the possible pairs of objects belonging to our test sources. After this, we asked the human expert to examine all the possible pairs of our test sources and to indicate us the semantic relationships that, in his opinion, existed among the corresponding objects.

At this point, we were able to evaluate the correctness and the completeness of our approach by measuring the classical parameters adopted in the literature for this purpose, i.e., Precision, Recall, F-Measure and Overall [441].

Precision is a measure of correctness. It is defined as:

$$Precision = \frac{|TP|}{|TP|+|FP|}$$

where TP are the true positives (i.e., semantic relationships detected by our approach and confirmed by the human expert), whereas FP are the false positives (i.e., semantic relationships proposed by our approach but not confirmed by our expert).

Recall is a measure of completeness. It is defined as:

$$Recall = \frac{|TP|}{|TP|+|FN|}$$

where FN are the false negatives (i.e., semantic relationships detected by the human expert that our system was unable to find).

F-Measure is the harmonic mean of Precision and Recall. It is defined as:

$$F\text{-}Measure = 2 \cdot \frac{Pecision \cdot Recall}{Precision + Recall}$$

Overall measures the post-match effort needed for adding false negatives and removing false positives from the set of matchings returned by the system to evaluate. It is defined as:

$$Overall = Recall \cdot (2 - \frac{1}{Precision})$$

Precision, Recall and F-Measure fall within the interval $[0, 1]$, whereas Overall ranges between $-\infty$ and 1; the higher Precision, Recall, F-Measure and Overall, the better the performance of the evaluated approach.

In Table 6.5, we report obtained results. From the analysis of this table, we can observe that, although our approach has been designed with the intent of privileging quickness and lightweightness over accuracy, for the reasons explained in the Introduction, its performance, in terms of correctness and completeness, is extremely satisfying.

We also point out that the values reported in Table 6.5 are those obtained by applying the threshold values reported in Section 6.5. These are the ones guaranteeing the best tradeoff between Precision and Recall and, consequently, the best values of F-Measure and Overall.

Interestingly, if, in a given application context, a user must privilege correctness (resp., completeness) over completeness (resp., correctness), it is sufficient to increase (resp., decrease) the values of th_{min} and to decrease (resp., increase) the values of th_{Ov} and th_{max}.

6.6.2 Evaluation of the pros and the cons of our approach

In order to provide a quantitative evaluation of the pros and the cons of our interschema property extraction approach w.r.t. the past ones thought for structured and

Property	*Precision*	*Recall*	*F-Measure*	*Overall*
Synonymies	0.82	0.87	0.84	0.68
Overlappings	0.77	0.69	0.73	0.48
Type Conflicts	0.78	0.73	0.75	0.52
Homonymies	0.95	0.92	0.93	0.87

Table 6.5. Precision, Recall, F-Measure and Overall of our approach

semi-structured sources[6] [373, 62], we compared our approach with XIKE [124]. Indeed, in [124], XIKE was already compared with several other systems having the same purposes (namely, Autoplex, COMA, Cupid, LSD, GLUE, SemInt, Similarity Flooding) and it was shown that it obtained comparable or better results.

First, we evaluated Precision, Recall, F-Measure and Overall of our approach and XIKE. Clearly, since this last system (as well as all the other ones mentioned above) did not handle unstructured data sources, we had to limit ourselves to consider only structured or semi-structured sources. Furthermore, as performed in [124], we limited our attention to synonymies and homonymies.

In a first experiment, we considered the same sources adopted in [124] for evaluating the performance of XIKE. In particular, we considered sources relative to Biomedical Data, Project Management, Property Register, Industrial Companies, Universities, Airlines, Biological Data and Scientific Publications. According to what reported in [124], Biomedical Schemas have been derived from several sites; among them we cite `http://www.biomediator.org`. Project Management, Property Register and Industrial Companies Schemas have been derived from Italian Central Governmental Office (ICGO) sources and are shown at the address `http://www.mat.unical.it/terracina/tests.html`. Universities Schemas have been downloaded from the address `http://anhai.cs.uiuc.edu/archive/domains/ courses.html`. Airlines Schemas have been found in [356]; Biological Schemas have been downloaded from the addresses `http://smi-web.stanford.edu/projects/helix/pubs/ismb02/schemas/`,
`http://www.cs.toronto.edu/db/clio/data/GeneX\_RDB-s.xsd` and `http://www.genome.ad.jp/`
`kegg/soap/v3.0/KEGG.wsdl`. Finally, Scientific Publications Schemas have been supplied by the authors of [256].

[6] Actually, to the best of our knowledge, no approach to uniformly extract interschema properties from structured, semi-structured and unstructured sources have been proposed in the past.

Application context	*Number of Schemas*	*Max depth*	*Average Number of nodes*	*Average Number of complex elements*
Biomedical Data	11	8	26	8
Project Management	3	4	40	7
Property Register	2	4	70	14
Industrial Companies	5	4	28	8
Universities	5	5	17	4
Airlines	2	4	13	4
Biological Data	5	8	327	60
Scientific Publications	2	6	18	9

Table 6.6. Characteristics of the sources adopted for evaluating our approach

System	*Precision*	*Recall*	*F-Measure*	*Overall*
XIKE (Synonymies)	0.92	0.90	0.91	0.82
XIKE (Homonymies)	0.87	0.95	0.91	0.81
Our approach (Synonymies)	0.84	0.87	0.85	0.70
Our approach (Homonymies)	0.79	0.92	0.85	0.68

Table 6.7. Precision, Recall, F-Measure and Overall of XIKE and our approach

We considered 35 sources whose characteristics are reported in Table 6.6. The minimum, the maximum and the average number of concepts of our sources were 12, 829 and 79, respectively.

The number of synonymies (resp., homonymies) really present in these sources was 498 (resp, 66). The number of synonymies (resp., homonymies) returned by XIKE was 541 (resp, 76). Finally, the number of synonymies (resp., homonymies) returned by our system was 593 (resp., 84). By comparing real synonymies and homonymies with the ones returned by XIKE and our approach we computed Precision, Recall, F-Measure and Overall for these two systems. They are reported in Table 6.7.

From the analysis of this table we can observe that Precision, Recall, F-Measure and Overall are better in XIKE, even if those obtained by our approach are satisfying. This was expected because our approach has been designed to be lightweight and, therefore, it introduces some approximations. For instance, while XIKE considers the neighbors of many levels in the computation of the similarity degree of two objects, our approach considers only the neighbors of levels 1 and 2.

Until now, our experimental campaign highlighted the cons of our approach. To evidence and quantify the pros, we measured its response time and the one of XIKE when the number of involved concepts represented in the corresponding metadata to examine increases. Obtained results are reported in Figure 6.6.

From the analysis of this figure, it clearly emerges that, as for this aspect, our approach is much better than XIKE. Indeed, the difference in the computation time

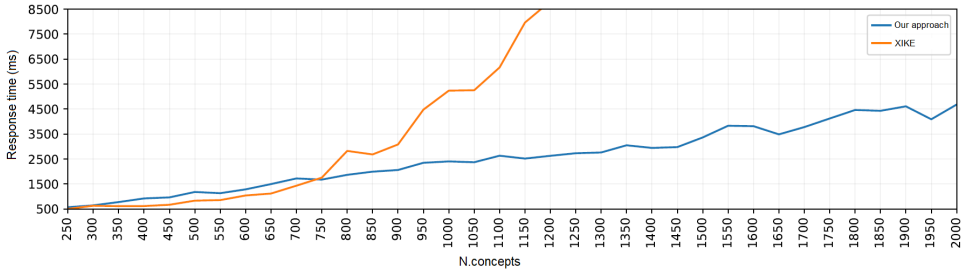


Fig. 6.6. Computation time of XIKE and our approach against the number of concepts to process

between it and XIKE is of various orders of magnitude and is such to make XIKE, and the other systems mentioned above, unsuitable to handle the number and the size of the data sources characterizing the current big data scenario.

With reference to this claim, we observe that, in this experiment, the response time is measured against the number of concepts in the source metaschema. As such, already a set of sources with 1500 concepts can be considered "large". Indeed, it would correspond, for instance, to a set of E/R schemas consisting of about 1500 entities or a set of XML Schemas defining about 1500 different element types.

Furthermore, in this analysis, we must not forget that XIKE and the approaches mentioned above are not capable of handling unstructured data, which represents the second (and, for many verses, most important) peculiarity of our approach.

6.6.3 A deeper investigation on the scalability of our approach

The previous experiment represents a first confirmation of the quickness and the scalability of our approach. In this section, we aim at finding a further confirmation of this trend by considering a much more numerous and articulated set of sources and by comparing the accuracy and the response time of our approach, of XIKE [124] and DIKE [347]. This last is one of the approaches of its generation showing the highest accuracy, as witnessed by the comparison tests described in [373].

Clearly, if we want to compare these three approaches, the only way of proceeding is to consider structured sources because they are the only ones handled by DIKE. In particular, we considered the database schemas of Italian Central Government Offices (hereafter, ICGO). They include about 300 databases that are heterogeneous both in the data model and languages (e.g., hierarchical, network, relational), as well as in their structure and complexity, ranging from simple databases, with schemas including few objects, to very complex databases [349].

Observe that our approach, XIKE and DIKE are all based on graphs and on the computation of similarities of the neighbors of the involved objects. However, DIKE was thought for relatively small structured databases. As a consequence, when it

System	*Precision*	*Recall*	*F-Measure*	*Overall*
DIKE (Synonymies)	0.98	0.93	0.95	0.91
DIKE (Homonymies)	0.96	0.95	0.95	0.91
XIKE $u = 5$ (Synonymies)	0.96	0.91	0.93	0.87
XIKE $u = 5$ (Homonymies)	0.93	0.93	0.93	0.86
XIKE $u = 2$ (Synonymies)	0.84	0.86	0.85	0.70
XIKE $u = 2$ (Homonymies)	0.85	0.86	0.85	0.71
Our approach (Synonymies)	0.83	0.81	0.82	0.64
Our approach (Homonymies)	0.81	0.83	0.82	0.64

Table 6.8. Precision, Recall, F-Measure and Overall of DIKE, XIKE ($u = 5$, $u = 2$) and our approach

computes the similarity of two objects belonging to different sources, it considers the similarity of their direct neighbors, the one of the neighbors of their direct neighbors, and so forth, until it terminates a fixpoint computation. In the worst case, the number of iterations of the fixpoint computation could be equal to the number of concepts of one of the involved sources. Clearly, performing such a high number of iterations allows DIKE to return very accurate results, but the required computation time is generally very high not only from the worst case computational complexity viewpoint, but also from the real computation time point of view. In XIKE, the possible number and dimension of data sources is higher than DIKE and they can be both structured and semi-structured. As a consequence, there is the need to limit the number of iterations of the fixpoint computation. For this reason, the concept of severity level is introduced. In particular, a user can choose a severity level u between 1 and n and the fixpoint computation is not completed but terminates after u iterations. The higher u the more accurate and slower XIKE. Our approach privileges lightweightness over accuracy for the reasons explained above. As a consequence, in this case, we limited the fixpoint computation to only 2 iterations. This could cause a reduction of accuracy but it is the only way to extend the approach of DIKE and XIKE also to a big data scenario.

Analogously to what happened in the previous section, in order to verify the theoretical conjectures explained above, we applied our approach, DIKE and XIKE (with $u = 5$ and, then, with $u = 2$) to ICGO databases. The obtained results are reported in Table 6.8.

The results of this table confirm our conjectures. DIKE provides a higher Precision, Recall, F-Measure and Overall than XIKE which, in turn, provides better results than our approach. Finally, XIKE, with a severity level equal to 5, provides better results than XIKE with a severity level equal to 2. The former tend to be comparable with

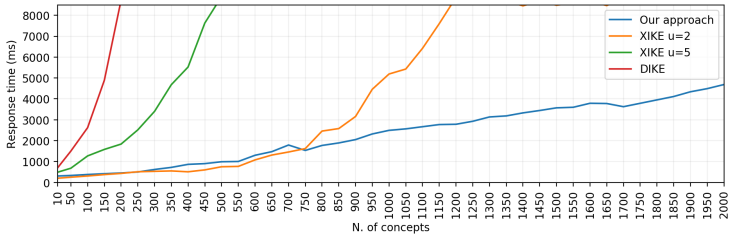


Fig. 6.7. Computation time of DIKE, XIKE ($u = 5$ and $u = 2$) and our approach against the number of concepts to process

the ones of DIKE; the latter tend to be comparable with the ones of our approach. This is in line with the fact that, when u tends to 5 the fixpoint computation tends to be complete; instead, when $u = 2$, it is substituted by only three iterations.

In any case, we would like to remark that, analogously to what happened in the previous experiment, the results obtained by our approach are still acceptable.

After having verified our conjectures about accuracy, we analyzed the ones regarding computation time. In particular, the average computation time of DIKE, XIKE (with $u = 5$ and $u = 2$) and our approach is reported in Figure 6.7.

From the analysis of this figure, it is easy to observe that the lower performance in terms of accuracy of our approach is largely balanced by an increased performance in terms of computation time. In a big data context, this aspect is mandatory. As a matter of fact, Figure 6.7 shows that DIKE and XIKE (especially when the severity level is high), even if very accurate, could not be applied in a big data scenario.

6.6.4 Evaluation of the role of our approach for structuring unstructured sources

As previously pointed out, one of the main contributions of this chapter is the approach for structuring unstructured sources. In the Introduction, we have seen that an important theoretical property of our approach (that distinguishes it from several possible alternative ones, like those based on ontologies) is its applicability to all possible scenarios, because it does not require a support knowledge, except for a (possibly generic) thesaurus, like BabelNet. In this section, we test its accuracy by comparing it with an alternative approach. For this purpose, we extended to unstructured data the clustering-based family of approaches defined for structured and semi-structured sources (see, for instance [27, 371]).

We performed this extension as follows: we considered the keywords associated with an unstructured source and used WordNet to derive a semantic distance coefficient for each pair of keywords. Then, we applied a clustering algorithm (specifically, Expectation Maximization [191]) to group keywords into homogeneous clusters. In

Property	*Precision*	*Recall*	*F-Measure*	*Overall*
Synonymies	0.76	0.82	0.79	0.56
Overlappings	0.69	0.65	0.67	0.36
Type Conflicts	0.72	0.64	0.68	0.39
Homonymies	0.91	0.88	0.89	0.79

Table 6.9. Precision, Recall, F-Measure and Overall of our approach when a clustering-based technique for structuring unstructured sources is applied

this way, we obtained a possible structure for unstructured sources. This structure is in line with what was done in the past for the clustering-based family of approaches, when they were applied on structured and semi-structured sources. This way of proceeding gave us the possibility to still apply the interschema property extraction approach defined in Section 6.5. In this case, we assumed that, given a keyword, the corresponding neighborhood consisted of the other keywords of its clusters.

We performed the same experiment described in Section 6.6.1 on the same sources. The only difference was the substitution of our approach for structuring unstructured sources with the clustering-based approach outlined above. The obtained results are shown in Table 6.9. Clearly, the differences between the performance reported in Tables 6.5 and 6.9 were due exclusively to the merits or demerits of our approach for structuring unstructured sources. From the analysis of this table we can observe that our approach presents a better performance than the corresponding clustering-based one described above. The differences are evident even if not extremely marked. For instance, we can observe a gain in Precision (resp., Recall, F-Measure, Overall) ranging from 4% (resp., 4%, 4%, 9%) to 10% (resp., 12%, 10%, 25%).

The results of this experiment, coupled with the theoretical analysis performed in the Introduction and mentioned above, allow us to conclude that our approach for structuring unstructured data is really capable of satisfying the requirements for which it was defined.

7

Extraction of Knowledge Patterns

7.1 Introduction

In the last few years, the "big data phenomenon" is rapidly changing the research and technological "coordinates" of the information system area [93, 451]. For instance, it is well known that data warehouses, generally handling structured and semi-structured data offline, are too complex and rigid to manage the wide amount and variety of rapidly evolving data sources of interest for a given organization, and the usage of more agile and flexible structures appears compulsory [128]. Data lakes are one of the most promising answers to this exigency. Differently from a data warehouse, a data lake uses a flat architecture (so that the insertion and the removal of a source can be easily performed). However, the agile and effective management of data stored therein is guaranteed by the presence of a rich set of extended metadata. These allow a very agile and easily configurable usage of the data stored in the data lake. For instance, if a given application requires the querying of some data sources, one could process available metadata to determine the portion of the involved data lake to examine.

One of the most radical changes caused by the big data phenomenon is the presence of a huge amount of unstructured data. As a matter of fact, it is esteemed that, currently, more than 80% of the information available on the Internet is unstructured [110]. In presence of unstructured data, all the approaches developed in the past for structured and semi-structured data must be "renewed", and the new approaches will be presumably much more complex than the old ones [228, 429]. Think, for instance, of schema integration: unstructured sources do not have a representing schema and, often, only a set of keywords are given (or can be extracted) to represent the corresponding content [129].

This chapter aims at providing a contribution in this setting. In particular, it proposes an approach to the extraction of complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. Here, we use the term "complex knowledge pattern" to indicate an intensional

relationship involving more concepts possibly belonging to different (and, presumably, heterogeneous) sources of a data lake. Formally speaking, in this chapter, a complex knowledge pattern consists of a logic succession $\{x_1, x_2, \ldots, x_w\}$ of w objects such that there is a semantic relationship (specifically, a synonymy or a part-of relationship) linking the k^{th} and the $(k+1)^{th}$ objects $(1 \leq k \leq w-1)$ of the succession.

Our approach is network-based in that it represents all the data lake sources by means of suitable networks. As a matter of facts, networks are very flexible structures, which allow the modeling of almost all phenomena that researchers aim at investigating. For instance, they have been used in the past to uniformly represent data sources characterized by heterogeneous, both structured and semi-structured, formats [124]. In this chapter, we also use networks to represent unstructured sources, which, as said before, do not have a representing schema. Furthermore, we propose a technique to construct a "structured representation" of the flat keywords generally used to represent unstructured data sources. This is a fundamental task because it highly facilitates the uniform management, through our network-based model, of structured, semi-structured and unstructured data sources.

Thanks to this uniform, network-based representation of the data lake sources, the extraction of complex knowledge patterns can be performed by exploiting graph-based tools. In particular, taking into consideration our definition of complex knowledge patterns, a possible approach for their derivation could consist in the construction of suitable paths going from the first node (i.e., x_1) to the last node (i.e., x_w) of the succession expressing the patterns. In this case, a user specifies the "seed objects" of the pattern (i.e., x_1 and x_w) and our approach finds a suitable path (if it exists) linking x_1 to x_w.

Since x_1 and x_w could belong to different sources, our approach should consider the possible presence of synonymies between concepts belonging to different sources, it should model these synonymies by means of a suitable form of arcs (cross arcs, or c-arcs), and should include both intra-source arcs (inner arcs, or i-arcs) and c-arcs in the path connecting x_1 to x_w and representing the complex knowledge pattern of interest.

Among all the possible paths connecting x_1 to x_w, our approach takes the shortest one (i.e., the one with the minimum number of c-arcs and, at the same number of c-arcs, the one with the minimum number of i-arcs). This choice is motivated by observing that a knowledge pattern should be as semantically homogeneous as possible. With this in mind, it is appropriate to reduce as much as possible the number of synonymies to be considered in the knowledge pattern from x_1 to x_w. This because a synonymy is weaker than an identity and, furthermore, it involves objects belonging to different sources which, inevitably, causes an "impairment" in the path going from

x_1 to x_w. Moreover, there is a further, more technological reason leading to the choice of the shortest path. Indeed, it is presumable that, after a complex knowledge pattern has been defined and validated at the intensional level, one would like to recover the corresponding data at the extensional level. In this case, in a big data scenario, reducing the number of the sources to query would generally reduce the volume and the variety of the data to process and, hence, would increase the velocity at which the data of interest are retrieved and processed.

As it will be clear in the following, there are cases in which synonymies (and, hence, c-arcs) are not sufficient to find a complex knowledge pattern from x_1 to x_w. In these cases, our approach performs two further attempts in which it tries to involve string similarities first, and, if even these properties are not sufficient, part-whole relationships. If neither synonymies nor string similarities nor part-whole relationships allow the construction of a path from x_1 to x_w, our approach concludes that, in the data lake into consideration, a complex knowledge pattern from x_1 to x_w does not exist.

Summarizing, the main contributions of this chapter are the following:

- it proposes a new network-based model to represent the structured, semi-structured and unstructured sources of a data lake;
- it proposes a new approach to, at least partially, "structuring" unstructured sources;
- it proposes a new approach to extracting complex knowledge patterns from the sources of a data lake.

This chapter is structured as follows: in Section 7.2, we illustrate related literature. In Section 7.3, we present our network-based model for data lakes. In Section 7.4, we describe our approach to enriching the representation of unstructured data sources in such a way as to, at least partially, "structure" them. In Section 7.5, we present our approach to the extraction of complex knowledge patterns. In Section 7.6, we describe some case studies conceived to illustrate the various possible behaviors of our approach. In Section 7.7, we present a critical discussion of several aspects concerning our approach.

7.2 Related Literature

In the literature there is a strong agreement in the definition of data lake. For instance, [188] defines data lakes as "big data repositories which store raw data and provide functionality for on-demand integration with the help of metadata descriptions". [434] claims that "a data lake is a set of centralized repositories containing vast amounts of raw data

(either structured or unstructured), described by metadata, organized into identifiable data sets, and available on demand". Analogously, [313] says that "a data lake refers to a massively scalable storage repository that holds a vast amount of raw data in its native format (≪as is≫) until it is needed plus processing systems (engine) that can ingest data without compromising the data structure". Finally, [375] says that "a data lake uses a flat architecture to store data in their raw format. Each data entity in the lake is associated with a unique identifier and a set of extended metadata, and consumers can use purpose-built schemas to query relevant data, which will result in a smaller set of data that can be analyzed to help answer a consumers question". A step forward, but in the same direction, can be found in [291], where the authors introduce the concept of Big Data Lake as "a central location in which users can store all their data in its native form, regardless of its source or format. Big data lake can be used as an environment for the development of in-depth analytics oriented toward fast decision making on the basis of raw data". Clearly, this strong agreement on the data lake definitions does not prevent the possibility to have very different architectures, management approaches and querying techniques in the data lake context, as we will see in the following.

The data lake paradigm requires each raw data to have associated a set of metadata. These represent a key component in the data lake architecture because they let data to be searchable and processed whenever this is necessary [447]. In [149], metadata are also used for bringing quality to a data lake. Here, the authors present CLAMS, a system for discovering integrity constraints from raw data and metadata. To validate obtained results, CLAMS needs human intervention.

In [151], the authors propose a data lake management approach that aims at extracting metadata from the Hive database. To reach its objective, it applies Social Network Analysis based techniques. Instead, in [411], iFuse, a data fusion platform that uses a Bayesian graphical model for both managing and querying a data lake, is proposed.

In the literature, there are many approaches to querying and managing both structured and semi-structured data (see [288, 60, 124, 349], to cite a few of them). However, they are generally incapable of managing unstructured data and are not lightweight and flexible enough to be used in the new data lake context. Furthermore, most of the approaches used for representing unstructured data are limited to texts [391]. In order to address this issue, the authors of [485, 96] propose a generalized data model to represent unstructured data, a method to process it (called RAISE) and an associated SQL-like query language. The authors of [155] propose the usage of machine learning for managing and extracting information from unstructured data. They motivate this proposal by observing that, currently, unstructured data represent about the 80% of

stored information and, therefore, they must be necessarily processed with a limited human intervention.

The extraction of Complex Knowledge Patterns (CKPs) is a topic widely investigated in the literature. This is due to the fact that CKPs can refer to a lot of research fields and, therefore, their extraction is a challenging issue in several research areas. Research concerning CKPs goes from keyword search and rank (see [116, 192, 174, 264], just to cite a few approaches) to visual knowledge extraction [388, 98]. In the literature, a huge variety of approaches to extracting CKPs has been proposed. Some of them are based on Network Analysis [473], others are centered on "questions and answers" mechanisms [219], further ones exploit Similarity Join [402], and so forth. Each family of approaches has its pros and cons, as well as its corresponding tools [407].

As for the approaches most related to ours, there are four main families that we need to investigate, namely: *(i)* extraction of keyword patterns; *(ii)* extraction of knowledge from structured sources; *(iii)* extraction of knowledge from heterogeneous sources; *(iv)* extraction of knowledge patterns through network analysis-based approaches.

As far as the first family is concerned, it is necessary to further differentiate the corresponding approaches. A first sub-family focuses on RDF analysis. In this context, several proposals can be found in the literature. For instance, in [116], approaches to keyword search inside RDF data are proposed. These approaches exploit user feedback to relax the search constraints and to identify a higher number of matches. The authors of [192] build a bipartite graph from RDF data and aim at solving a graph assembly problem. Since this problem is NP-hard, they propose two heuristics for facing it. In [324], models letting knowledge patterns to be represented by means of RDF are investigated. The second sub-family, instead, aims at extracting keyword patterns in a graph database. In [196], the authors propose BLINKS, a system consisting of an algorithm for bi-level indexing and a query processor useful for searching the top-k keywords in a graph. In [174], an engine for enumerating keywords and evaluating their search in a data graph is proposed. In the same way, in [264], EASE, a framework allowing indexing and keyword querying, is described.

As for the second family, most of the corresponding approaches are summarized in [278]. Here, the authors claim that, thanks to metadata, it is possible to think of a new, completely automated, approach.

As for the third family, in [100], the authors provide an overview of techniques used in the literature to support keyword search in structured and semi-structured data. In [274], the authors operate on semi-structured sources and try to make the extraction process as automated as possible. More recent approaches try to extract knowledge

from heterogeneous sources. In fact, as evidenced in [271], the big data phenomenon led to the creation of a lot of heterogeneous sources that include unstructured data. These need to be integrated exactly as it was made before for structured data. Starting from this consideration, the authors analyze the most important challenges introduced by this new reality and present a unique query format taking this issue into account. In [388], the authors assert that, in order to have a user-friendly graph query engine, it is necessary to support different kinds of task, like synonymy detection and ontology usage. Based on this assertion, they propose a framework allowing these operations on data without schema or structure. In [402], the authors argue that Similarity Join is a fundamental operation for clearing data and integrating different sources. It involves two big challenges, namely quantifying knowledge aware similarities and identifying similarity pairs efficiently. To address these issues, they propose a new framework. Likewise, in [431], a system to integrate different sources through keyword search, and an evaluation system based on user feedback, are proposed.

The last family of approaches is based on network analysis. In [393], network communities and the apriori algorithm are used to identify rhythmic knowledge patterns of musical work. In [281], the authors represent patent data as a network and, then, propose a new approach that analyzes this network for extracting CKPs about patent applicants. In [290], the authors propose a new formalism to represent a knowledge base through a network whose edges denote the semantic proximity between two or more concepts. This representation allows the discovery of association models among different concepts. In [227], the authors propose an algorithm that uses the cliques in a graph for searching the keywords linked to a given input. According to what the authors claim, keyword search is necessary because it facilitates the identification of sub-graphs in a network.

7.3 A network-based model for data lakes

In this section, we illustrate our network-based model to represent and handle a data lake, which we will use in the rest of this chapter.

In our model, a data lake DL is represented as a set of m data sources:

$$DL = \{D_1, D_2, \cdots, D_m\}$$

A data source $D_k \in DL$ is provided with a rich set $\mathcal{M}_k$ of metadata. We denote with $\mathcal{M}_{DL}$ the repository of the metadata of all the data sources of DL:

$$\mathcal{M}_{DL} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$$

According to [341], our model represents $\mathcal{M}_k$ by means of a triplet:

$$\mathcal{M}_k = \langle \mathcal{M}_k^T, \mathcal{M}_k^O, \mathcal{M}_k^B \rangle$$

Here:

- $\mathcal{M}_k^T$ denotes *technical metadata*. They represent the type, the format, the structure and the schema of the corresponding data. They are commonly provided by the source catalogue.
- $\mathcal{M}_k^O$ represents *operational metadata*. They include the source and target locations of the corresponding data, the associated file size, the number of their records, and so on. Usually, they are automatically generated by the technical framework handling the data lake.
- $\mathcal{M}_k^B$ indicates *business metadata*. They comprise the business names and descriptions assigned to data fields. They also cover business rules, which can become integrity constraints for the corresponding data source.

Since our approach focuses on the semantics of data sources, in this chapter, we consider only business metadata. Indeed, they denote, at the intensional level, the information content stored in $\mathcal{M}_k$ and are those of interest for supporting the extraction of complex knowledge patterns from a data lake, which is our ultimate goal.

Our model adopts a notation typical of XML, JSON and many other semi-structured models to represent $\mathcal{M}_k^B$. According to this notation, Obj_k indicates the set of all the objects stored in $\mathcal{M}_k^B$. It consists of the union of three subsets:

$$Obj_k = Att_k \cup Smp_k \cup Cmp_k$$

Here:

- Att_k indicates the set of the attributes of $\mathcal{M}_k^B$;
- Smp_k represents the set of the simple elements of $\mathcal{M}_k^B$;
- Cmp_k denotes the set of the complex elements of $\mathcal{M}_k^B$.

Here, the meaning of the terms "attribute", "simple element" and "complex element" is the one typical of semi-structured data models.

$\mathcal{M}_k^B$ can be also represented as a graph:

$$\mathcal{M}_k^B = \langle N_k, A_k \rangle$$

N_k is the set of the nodes of $\mathcal{M}_k^B$. There exists a node $n_{k_j} \in N_k$ for each object $o_{k_j} \in Obj_k$. According to the structure of Obj_k, N_k consists of the union of three subsets:

$$N_k = N_k^{Att} \cup N_k^{Smp} \cup N_k^{Cmp}$$

Here, N_k^{Att} (resp., N_k^{Smp}, N_k^{Cmp}) indicates the set of the nodes corresponding to Att_k (resp., Smp_k, Cmp_k). There is a biuniovocal correspondence between a node of N_k and an object of Obj_k. Therefore, in the following, we will use the two terms interchangeably.

Let x be a complex element of $\mathcal{M}_k^B$. Obj_{k_x} indicates the set of the objects directly contained in x, whereas $N_{k_x}^{Obj}$ denotes the set of the corresponding nodes. Furthermore, let x be a simple element of $\mathcal{M}_k^B$. Att_{k_x} represents the set of the attributes directly contained in x, whereas $N_{k_x}^{Att}$ denotes the set of the corresponding nodes.

A_k indicates the set of the arcs of $\mathcal{M}_k^B$. It consists of three subsets:

$$A_k = A'_k \cup A''_k \cup A'''_k$$

Here:

- $A'_k = \{(n_x, n_y) | n_x \in N_k^{Cmp}, n_y \in N_{n_x}^{Obj}\}$. This definition specifies that there is an arc from a complex element of $\mathcal{M}_k^B$ to each object directly contained in it.
- $A''_k = \{(n_x, n_y) | n_x \in N_k^{Smp}, n_y \in N_{n_x}^{Att}\}$. This definition specifies that there is an arc from a simple element of $\mathcal{M}_k^B$ to each attribute directly contained in it.
- $A'''_k = \{(n_x, n_y) | n_x \in N_k, n_y \in N_k$, D_k is unstructured and between n_x and $n_y\}$ there exists a correlation. The meaning of A'''_k will be clear after reading Section 7.4, where we illustrate our approach for "structuring" unstructured data.

Interestingly, our data lake formalization uses a model similar to the one adopted in [291]. Here, a data lake is defined as a pair $DL = \{V, M\}$, where V is a set of values in the data lake and M is a set of metadata describing the values of DL. In this definition, the authors introduce the concept of fully description in terms of attribute names and data types. This definition is similar to the components $\mathcal{M}_k^T$ and $\mathcal{M}_k^O$ of our model's metadata. However, the two approaches present several differences because our own also introduces the concept of business metadata (thus enriching the data description component), whereas the approach of [291] proceeds with the formal definition of the Extract, Process and Store (EPS) process (thus enriching the process description component).

7.4 Enriching the representation of unstructured data

Our network-based model for representing and handling a data lake is perfectly fitted for representing and managing semi-structured data because it has been designed having XML and JSON in mind. Clearly, it is sufficiently powerful to represent structured data. The highest difficulty regards unstructured data because it is worth avoiding a flat representation consisting of a simple element for each keyword provided to denote the source content. As a matter of fact, this kind of representation would make

the reconciliation, and the next integration, of an unstructured source with the other (semi-structured and structured) ones of the data lake very difficult. Therefore, it is necessary to (at least partially) "structure" unstructured data.

Our approach to addressing this issue creates a complex element for representing the source as a whole and a simple element for each keyword. Furthermore, it adds an arc from the source to each of the simple elements. Initially, there is no arc between two simple elements. To determine the arcs to add, our approach exploits lexical and string similarities.

In particular, lexical similarity is considered by stating that there exists an arc from the node n_{k_1}, corresponding to the keyword k_1, to the node n_{k_2}, corresponding to the keyword k_2 (and vice versa), if k_1 and k_2 have at least one common lemma[1] in a suitable thesaurus. Taking the current trends into account, this thesaurus should be a multimedia one; for this purpose, in our experiments, we have adopted BabelNet [326]. When this pair of arcs has been added, n_{k_1} and n_{k_2} must be considered complex elements, instead of simple elements.

String similarity is applied by stating that there exists an arc from n_{k_1} to n_{k_2} (and vice versa) if the string similarity degree $kd(k_1, k_2)$, computed by applying a suitable string similarity metric on k_1 and k_2, is "sufficiently high" (see below). We have chosen N-Grams [241] as string similarity metric because we have experimentally seen that it provides the best results in our context. Also in this case, when this pair of arcs has been added, n_{k_1} and n_{k_2} change their types from simple elements to complex ones. Now, we illustrate in detail what "sufficiently high" means and how our approach operates. Let $KeySim$ be the set of the string similarities for each pair of keywords of the source into consideration. Each record in $KeySim$ has the form $\langle k_i, k_j, kd(k_i, k_j)\rangle$. Our approach first computes the maximum keyword similarity degree kd_{max} present in $KeySim$. Then, it examines each keyword similarity registered therein. Let $\langle k_1, k_2, kd(k_1, k_2)\rangle$ be one of these similarities. If $((kd(k_1, k_2) \geq th_k \cdot kd_{max})\, and \quad (kd(k_1, k_2) \geq th_{kmin}))$, which implies that the keyword similarity degree between k_1 and k_2 is among the highest ones in $KeySim$ and that, in any case, it is higher than or equal to a minimum threshold, then an arc is added from n_{k_1} to n_{k_2}, and vice versa. We have experimentally set $th_k = 0.70$ and $th_{kmin} = 0.50$.

From this description, it emerges that, given two nodes n_{k_1} and n_{k_2}, corresponding to two keywords k_1 and k_2 of the unstructured source, four cases may exist, namely: *(1)* no arcs link n_{k_1} and n_{k_2}; *(2)* an arc derived from an object similarity links them;

[1] In this chapter, we use the term "lemma" according to the meaning it has in BabelNet [326]. Here, given a term, its lemmas are other objects (terms, emoticons, etc.) semantically associated with it and, therefore, contributing to specify its meaning.

(3) an arc derived from a string similarity links them; *(4)* an arc derived from both an object and a string similarity links them.

In our approach, the component devoted to "structuring" unstructured data, which we are describing in this section, plays a key role. On the other hand, this last issue has been investigated in the recent past. For instance, in Section 7.3, we have cited the approach of [291] and we have seen that an important component of this approach is the EPS (Extract, Process and Store) process. The management of unstructured data is performed during the extract subtask of this process, when data are extracted from the data lake. This last is represented as a pair consisting of values and metadata. Instead, dataset schemas are dynamically defined, according to the "schema on read" approach. Starting from these three elements, the approach of [291] generates a rowset with n attributes.

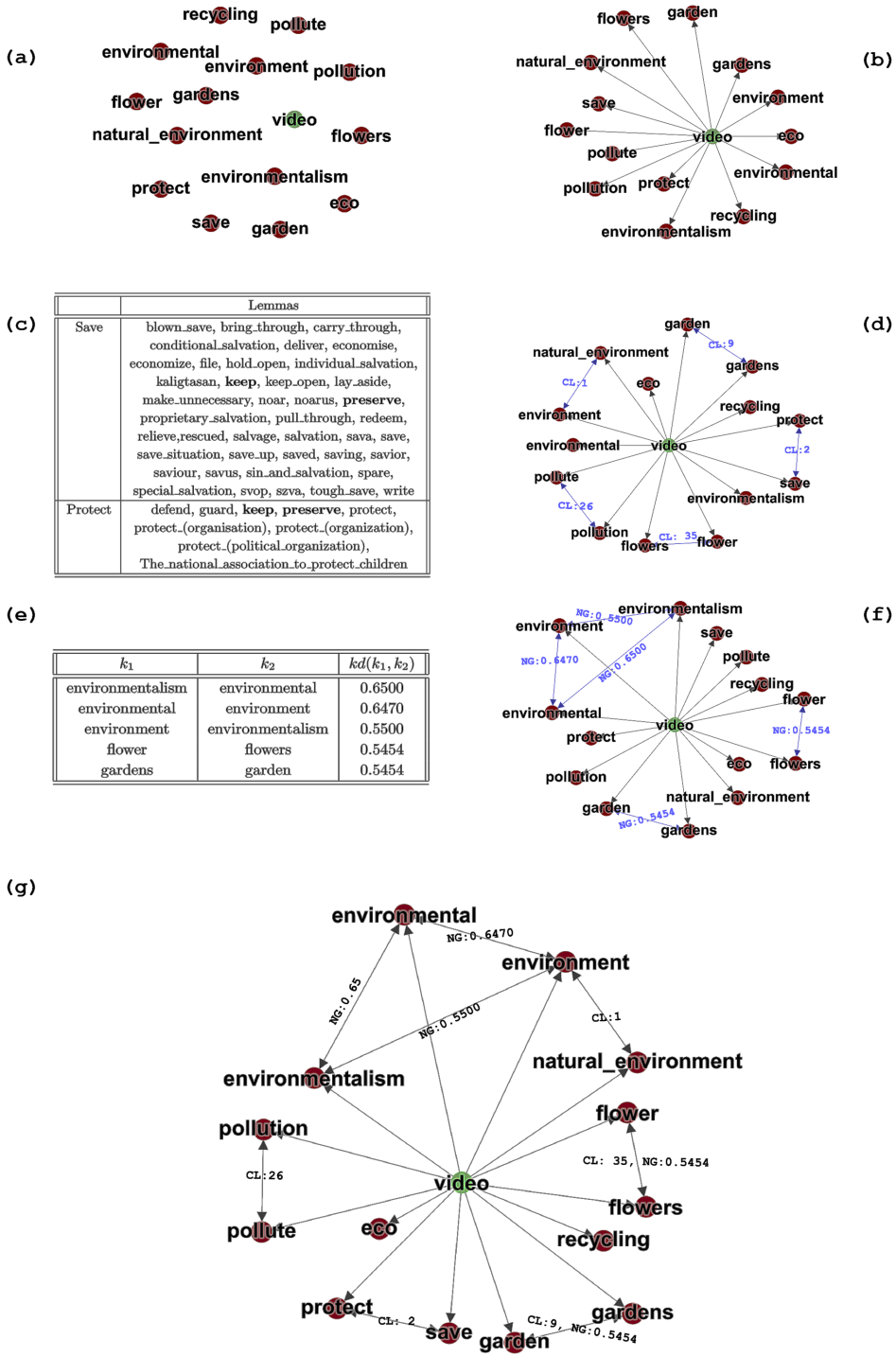
To correctly interpret data and/or metadata of unstructured sources, in the construction of rowset, the approach of [291] uses some transformation rules allowing the extraction and/or the correction of values and data acquired from the involved sources. To perform this task, the approach uses U-SQL and fuzzy logics. As a consequence of this way of proceeding, it can generate a tabular representation (i.e., the rowset) from unstructured data. The content of the rowset depends on the membership function associated with the fuzzy logic and on the possible constraints regarding it.

Our approach operates in a different way. Indeed, to perform structuring of unstructured sources, it leverages network analysis, as well as lexical and string similarities. In fact, unstructured sources are "structured" thanks to the addition of the arcs in the networks representing the sources themselves. These arcs can be created only when the similarity between nodes is higher than a certain degree. Interestingly, in both approaches, the final result of the structuring activity depends on a threshold.

The approach of [291] addresses the data variety issue by extending the operations that can be performed on unstructured sources by means of fuzzy techniques. These carry out the structuring task, and the consequent rowset creation, by means of an interface for the dataset extraction, which is unified and valid for all the sources. By contrast, our approach bases the structuring activity on business constraints involving the schemas of the data lake sources and on lexical and string similarities among the elements represented therein.

7.4.1 Example

Consider an unstructured data source consisting of a video about environment and pollution. As we said before, for each unstructured source, our approach begins from a set of keywords representing its content. In order to keep our description simple and



	Lemmas
Save	blown_save, bring_through, carry_through, conditional_salvation, deliver, economise, economize, file, hold_open, individual_salvation, kaligtasan, **keep**, keep_open, lay_aside, make_unnecessary, noar, noarus, **preserve**, proprietary_salvation, pull_through, redeem, relieve,rescued, salvage, salvation, sava, save, save_situation, save_up, saved, saving, savior, saviour, savus, sin_and_salvation, spare, special_salvation, svop, szva, tough_save, write
Protect	defend, guard, **keep**, **preserve**, protect, protect_(organisation), protect_(organization), protect_(political_organization), The_national_association_to_protect_children

k_1	k_2	$kd(k_1, k_2)$
environmentalism	environmental	0.6500
environmental	environment	0.6470
environment	environmentalism	0.5500
flower	flowers	0.5454
gardens	garden	0.5454

Fig. 7.1. Graphical representation of our approach to deriving a "structure" for an unstructured source

clear, in this example, we assume that our video has a limited number of keywords, namely the ones shown in Figure 6.1.

First, as we can see in Figure 6.1(a), our approach constructs a graph having a node for each keyword. A further node is added to represent the video as a whole;

nodes corresponding to keywords are colored in red, whereas the other one is colored in green. Following our strategy, in Figure 6.1(b), we added an arc from the node representing the whole video to each node associated with a keyword. The next step consists of using BabelNet. In Figure 6.1(c), we show two keywords ("Save" and "Protect") and the corresponding lemmas in BabelNet. Common lemmas (i.e., "keep" and "preserve") are in bold. Since "Save" and "Protect" have at least one common lemma, two arcs are added between the corresponding nodes in Figure 6.1(d). These arcs are highlighted in blue in this figure and, due to layout reasons, we report only one arc with two arrows, instead of two arcs with one arrow. Each arc has a label representing the number of common lemmas between the corresponding keywords in BabelNet. After having added the new arcs, caused by the common lemmas present in BabelNet, we proceed by analyzing string similarities. In Figure 6.1(e), we report the pairs of keywords that satisfy this feature. In Figure 6.1(f), we add a pair of arcs for each pair of keywords of Figure 6.1(e). Again, these arcs are highlighted in blue and, due to layout reasons, we report only one arc with two arrows, instead of two arcs with one arrow. Each arc has a label representing the string similarity degree (computed by means of N-Grams) between the corresponding keywords. Finally, in Figure 6.1(g), we combine the arcs derived in the previous two steps. Clearly, it may happen that, for a pair of keywords (see, for instance, the keywords "garden" and "gardens"), two pairs of arcs have been generated, one in Figure 6.1(d) and one in Figure 6.1(f). In this case, in Figure 6.1(g), we do not report two pairs of arcs; instead, we report only one pair, representing both of them. The label of this pair is obtained by merging the labels of the two corresponding pairs.

7.5 Extraction of complex knowledge patterns

7.5.1 General description of the approach

Our approach to the extraction of complex knowledge patterns operates on a data lake DL whose data sources are represented by means of the formalism described in Section 7.4.

It receives a dictionary Syn of synonymies involving the objects stored in the sources of DL. This dictionary could be a generic thesaurus, such as BabelNet [326], a domain-specific thesaurus, or a dictionary obtained by taking into account the structure and the semantics of the sources, which the corresponding objects refer to (such as the dictionaries produced by XIKE [124], MOMIS [60] or Cupid [288]). Furthermore, it receives two objects x_{i_h}, belonging to a source D_h of DL, and x_{j_q}, belonging to a source D_q of DL. x_{i_h} and x_{j_q} represent the base on which the complex knowledge pattern to extract should be constructed. As a matter of fact, we recall that, in this

chapter, a complex knowledge pattern consists of a logic succession $\{x_1, x_2, \ldots, x_w\}$ of w objects such that: *(i)* x_1 coincides with x_{i_h}; *(ii)* x_w coincides with x_{j_q}; *(iii)* there is a semantic relationship (specifically, a synonymy or a part-of relationship) linking the k^{th} and the $(k+1)^{th}$ objects $(1 \leq k \leq w-1)$ of the succession.

Before illustrating our approach in detail, it is necessary to introduce some preliminary concepts, namely the ones of i-arcs, c-arcs and pw-arcs. In Section 7.4, we have seen that, given a source D_k of DL, $\mathcal{M}_B^k = \langle N_k, A_k \rangle$ and $A_k = A'_k \cup A''_k \cup A'''_k$. All the arcs of A_k are internal to D_k; we call them *i-arcs* (i.e., internal arcs) in the following. Now, let x_{i_h} and x_{j_q} be two objects for which a synonymy exists in Syn. We call *c-arcs* (i.e., cross arcs) the arcs (n_{i_h}, n_{j_q}) and (n_{j_q}, n_{i_h}) corresponding to this synonymy. These arcs are extremely important because they allow the extraction, the processing and the management of information coming from different sources. Finally, given an arc $(n_{i_k}, n_{j_k}) \in A'_k \cup A''_k$, we call *pw-arc* (i.e., part-whole arc) the arc (n_{j_k}, n_{i_k}). The pw-arc (n_{j_k}, n_{i_k}) is the "reverse" arc of (n_{i_k}, n_{j_k}) because it starts from the part and ends to the whole[2]. The name of this arc clearly indicates its nature. As it is evident from the definition of A'_k and A''_k, the i-arc (n_{i_k}, n_{j_k}) specifies the existence of a part-of relationship, from the whole (n_{i_k}) to the part (n_{j_k}). The arc (n_{j_k}, n_{i_k}) is the reverse one going from the part to the whole.

Our approach operates as follows. Let n_{i_h} (resp., n_{j_q}) be the node corresponding to x_{i_h} (resp., x_{j_q}).

- If $h = q$, we have a trivial case and the complex knowledge pattern from n_{i_h} to n_{j_q} is obtained by selecting the set of the arcs belonging to the shortest path from n_{i_h} to n_{j_q}.
- If $h \neq q$, then c-arcs and pw-arcs must be considered. First, our approach determines the set of complex knowledge patterns each formed by a c-arc from n_{i_h} to a node n_{t_l} synonymous of n_{i_h}, followed by a complex knowledge pattern from n_{t_l} to n_{j_q}. Then, it determines the set of complex knowledge patterns each formed by an i-arc from n_{i_h} to a node n_{t_h}, being a part of n_{i_h}, followed by a complex knowledge pattern from n_{t_h} to n_{j_q}. If at least one of these two sets is not empty, it returns the complex knowledge pattern having the minimum number of c-arcs.

If both these sets are empty, then our approach performs a last attempt to find a complex knowledge pattern by considering pw-arcs having n_{i_h} as target, if they exist. In this case, it determines the set of complex knowledge patterns each formed by a pw-arc from n_{i_h} to a node n_{t_h} followed by a complex knowledge pattern from n_{t_h}

[2] In order to keep the layout simple, in the graphical representation of our model, we explicitly show only the arcs from the whole to the parts; the corresponding pw-arcs are considered implicit.

to n_{j_q}. If this set is not empty, it returns the complex knowledge pattern having the minimum number of pw-arcs.

7.5.2 Technical Details

As previously pointed out, our approach operates on a data lake DL. It receives a dictionary Syn of synonymies regarding the objects of DL, along with two objects x_{i_h}, belonging to a source D_h of DL, and x_{j_q}, belonging to a source D_q of DL. Let n_{i_h} (resp., n_{j_q}) be the node corresponding to x_{i_h} (resp., x_{j_q}), then the computation of $CKP(n_{i_h}, n_{j_q})$, i.e. of the complex knowledge pattern from n_{i_h} to n_{j_q}, is recursively performed as follows:

- If $h = q$, x_{i_h} and x_{j_q} belong to the same source and, therefore, n_{i_h} and n_{j_q} belong to the same network. In this case, the complex knowledge pattern $CKP(n_{i_h}, n_{j_q})$ from n_{i_h} to n_{j_q} is obtained by selecting the set of the arcs belonging to the shortest path from n_{i_h} to n_{j_q}. Any algorithm previously proposed in the literature for computing the shortest path between two nodes can be adopted.
- If $h \neq q$, then n_{i_h} and n_{j_q} belong to different networks.
  First, the set $NSynSet_{i_h}$ of the nodes corresponding to the objects synonymous of x_{i_h} in Syn is determined as:

$$NSynSet_{i_h} = \{n_{t_l} \mid (n_{i_h}, n_{t_l}) \in Syn\}$$

  Then, the set $CKPSynSet(n_{i_h}, n_{j_q})$ of the complex knowledge patterns from n_{i_h} to n_{j_q} and passing through a node of $NSynSet_{i_h}$ is computed. Formally:

$$CKPSynSet(n_{i_h}, n_{j_q}) = \{SynCKP(n_{i_h}, n_{j_q}, n_{t_l}) | n_{t_l} \in NSynSet_{i_h}\}$$

  where:

$$SynCKP(n_{i_h}, n_{j_q}, n_{t_l}) = \{(n_{i_h}, n_{t_l}) \cup CKP(n_{t_l}, n_{j_q})\}$$

  After this, the set $NPartSet_{i_h}$ of the nodes representing a part of n_{i_h} (which, in this case, is the whole) is determined as:

$$NPartSet_{i_h} = \{n_{t_h} | (n_{i_h}, n_{t_h}) \in A'_h \cup A''_h\}$$

  Then, the set $CKPPartSet(n_{i_h}, n_{j_q})$ of the complex knowledge patterns from n_{i_h} to n_{j_q} and passing through a node of $NPartSet_{i_h}$ is computed. Formally:

$$CKPPartSet(n_{i_h}, n_{j_q}) = \{PartCKP(n_{i_h}, n_{j_q}, n_{t_h}) | n_{t_h} \in NPartSet_{i_h}\}$$

  where:

$$PartCKP(n_{i_h}, n_{j_q}, n_{t_h}) = \{(n_{i_h}, n_{t_h}) \cup CKP(n_{t_h}, n_{j_q})\}$$

If $CKPSynSet(n_{i_h}, n_{j_q}) \neq \emptyset$ and/or $CKPPartSet(n_{i_h}, n_{j_q}) \neq \emptyset$, our approach selects as $CKP(n_{i_h}, n_{j_q})$ the complex knowledge pattern having the minimum number of c-arcs. If more than one pattern exists with the same minimum number of c-arcs, it returns the one with the minimum number of i-arcs. If more than one pattern exists with these characteristics, it randomly selects one of them.
If $CKPSynSet(n_{i_h}, n_{j_q}) = \emptyset$ and $CKPPartSet(n_{i_h}, n_{j_q}) = \emptyset$, then c-arcs are not sufficient to find a complex knowledge pattern from n_{i_h} to n_{j_q}. However, a last attempt to find such a pattern can be performed by exploiting a pw-arc having n_{i_h} as target, if it exists.
In particular, let $NWholeSet_{i_h}$ be the set of the nodes of which n_{i_h} is a part. It is determined as:

$$NWholeSet_{i_h} = \{n_{t_h} | (n_{t_h}, n_{i_h}) \in A'_h \cup A''_h\}$$

Then, if $NWholeSet_{i_h} = \emptyset$, there is no possibility to find a complex knowledge pattern from n_{i_h} to n_{j_q}. Otherwise, the set $CKPWholeSet(n_{i_h}, n_{j_q})$ of the complex knowledge patterns between n_{i_h} and n_{j_q} and passing through a node of $NWholeSet_{i_h}$ is computed. Formally:

$$CKPWholeSet(n_{i_h}, n_{j_q}) = \{WholeCKP(n_{i_h}, n_{j_q}, n_{t_h}) | n_{t_h} \in NWholeSet_{i_h}\}$$

where:

$$WholeCKP(n_{i_h}, n_{j_q}, n_{t_h}) = \{(n_{i_h}, n_{t_h}) \cup CKP(n_{t_h}, n_{j_q})\}$$

Once $WholeCKP(n_{i_h}, n_{j_q}, n_{t_h})$ has been constructed, if it is not empty, our approach selects as $CKP(n_{i_h}, n_{j_q})$ the complex knowledge pattern having the minimum number of pw-arcs. If more than one pattern exists with the same minimum number of pw-arcs, it returns the one with the minimum number of c-arcs. If more than one pattern exists with these characteristics, it selects the one with the minimum number of i-arcs. Finally, if more than one pattern exists with the same minimum number of i-arcs, it randomly selects one of them.

Computational complexity

As for the computational complexity of this approach, we can observe that:

- If $h = q$, any algorithm previously proposed in the literature for computing the shortest path between two nodes can be adopted. For instance, if the Dijkstra algorithm using a binary heap is implemented, the computational complexity of this step is $O(|A| \cdot log|N|)$, where $|A|$ is the total number of arcs of the data lake and $|N|$ is the total number of its nodes.

- If $h \neq q$, in the worst case, it is necessary to determine the sets $NSynSet_{i_h}$, $NPartSet_{i_h}$ and $NWholeSet_{i_h}$ and, then, for each node of these sets, to compute the shortest path from n_i to n_j bounded to pass through it.

  Now, $|NSynSet_{i_h}|$ is $O(|DL|)$ because there could be at most one synonymous of a node for each source. $|NPartSet_{i_h}|$ is $O(|N_{max}|)$, where $|N_{max}|$ is the number of nodes of the largest source of the data lake. For the same reason, $|NWholeSet_{i_h}|$ is $O(|N_{max}|)$.

  The complexity of the computation of the shortest path from n_i to n_j bounded to pass through a node is $O(|A| \cdot log(|N|)$, if the Dijkstra algorithm with the support of the binary heap is applied.

  Therefore, in this case, the computational complexity of the algorithm is:

$$O(|A| \cdot log|N|) \cdot O(max(|N_{max}|, |DL|))$$

  Now, since generally $|N_{max}| \gg |DL|$, we have that the computational complexity of this step is:

$$O(|A| \cdot log|N| \cdot |N_{max}|)$$

Since the computational complexity of the case $h \neq q$ is higher than the one of the case $h = q$, we can conclude that the overall computational complexity of our approach is $O(|A| \cdot log|N| \cdot |N_{max}|)$.

This computational complexity can be judged very good if we consider the problem to solve. Actually, one could say that it is high for real data lakes consisting of many sources. However, in these cases, we have that, in the reality, the corresponding graphs are very sparse and, therefore, $|A|$ is small. To better quantify this fact, in Section 7.7.3, we compare the theoretical and the real computation time of our approach against the number of nodes of the data lake.

7.6 Some case studies

In this section, we present some case studies devoted to illustrate the behavior of our approach in the various possible cases. To perform our test cases, we constructed a data lake consisting of 2 structured sources, 4 semi-structured sources (i.e., 2 XML sources and 2 JSON ones) and 4 unstructured sources (i.e., 2 books and 2 videos). All these sources store data about environment and pollution. To describe unstructured sources, we initially considered a set of keywords derived from Google Books, for books, and from YouTube, for videos. The interested reader can find the schemas, in case of structured and semi-structured sources, and the keywords, in case of unstructured sources, at the address `http://www.barbiana20.unirc.it/dls/datasets/dl2`. The password to type is "`za.12&;lq74:#`".

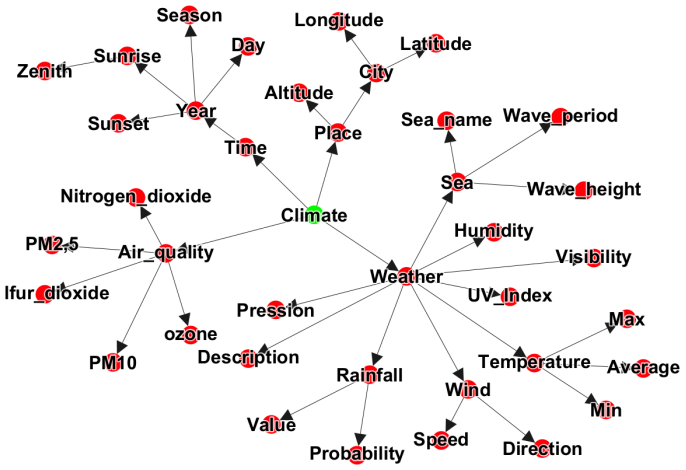


Fig. 7.2. The network corresponding to the source *Climate*

The case studies we present in this section involve five sources of our data lake. These sources are the following:

- *Climate*. This is a JSON source storing data about weather and climatic conditions of several places. In this dataset, space and time are expressed at several granularity levels. In particular, time is expressed in years, seasons and days, whereas space is expressed in countries and cities; these last ones have associated the corresponding latitude and longitude. The network representing *Climate* is reported in Figure 7.2.
- *Energy*. This is a JSON source storing data about renewable and non-renewable energy sources used in the countries worldwide. Energy also stores data about the investments on the various kinds of energy source. The network representing *Energy* is reported in Figure 7.3.
- *Environment disasters*. This is an XML source storing data about environment disasters (e.g., earthquakes, seaquakes, volcanic eruptions, etc.), the places where they happened, the damages caused by them, and so forth. The network representing *Environment disasters* is reported in Figure 7.4.
- *Environment risks*. This is a book discussing about environment risks, their probabilities, their damages, etc. This is an unstructured source and, as such, it is represented by a set of keywords, which is reported in Table 7.1.
- *Air pollution*. This is a book focusing on air pollution, its causes, its consequences and the possible control strategies that can mitigate their impact on the environment. This is an unstructured source. Its keywords are reported in Table 7.2.

The first case study we are considering is very simple because the two "seed objects" of the complex knowledge pattern are `Energy` and `Population`, both belonging to the source *Energy*. Since both the source and the target node of the knowledge

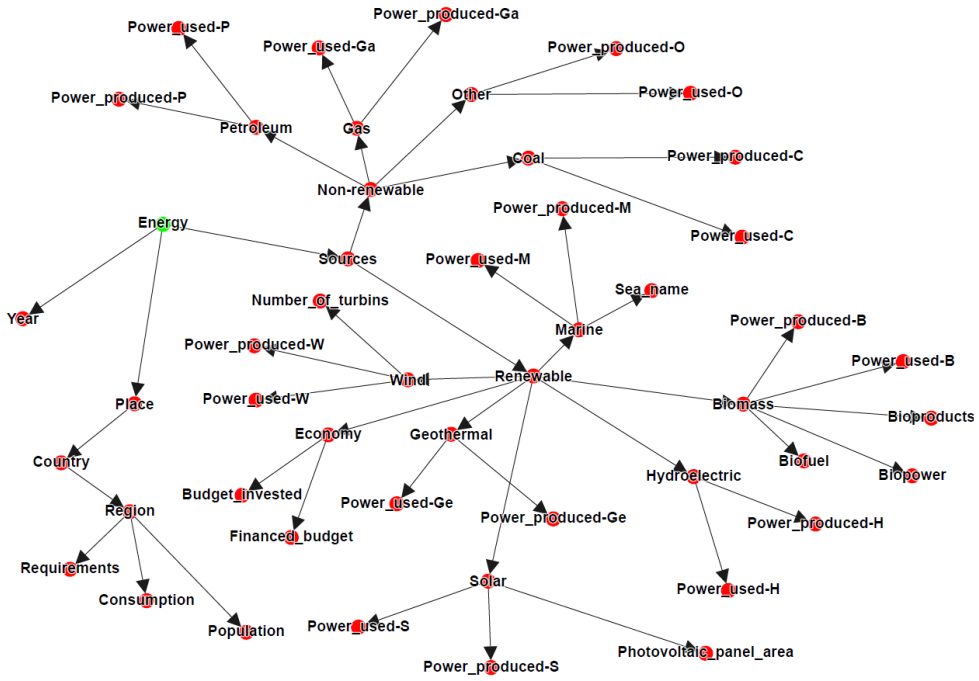


Fig. 7.3. The network corresponding to the source *Energy*

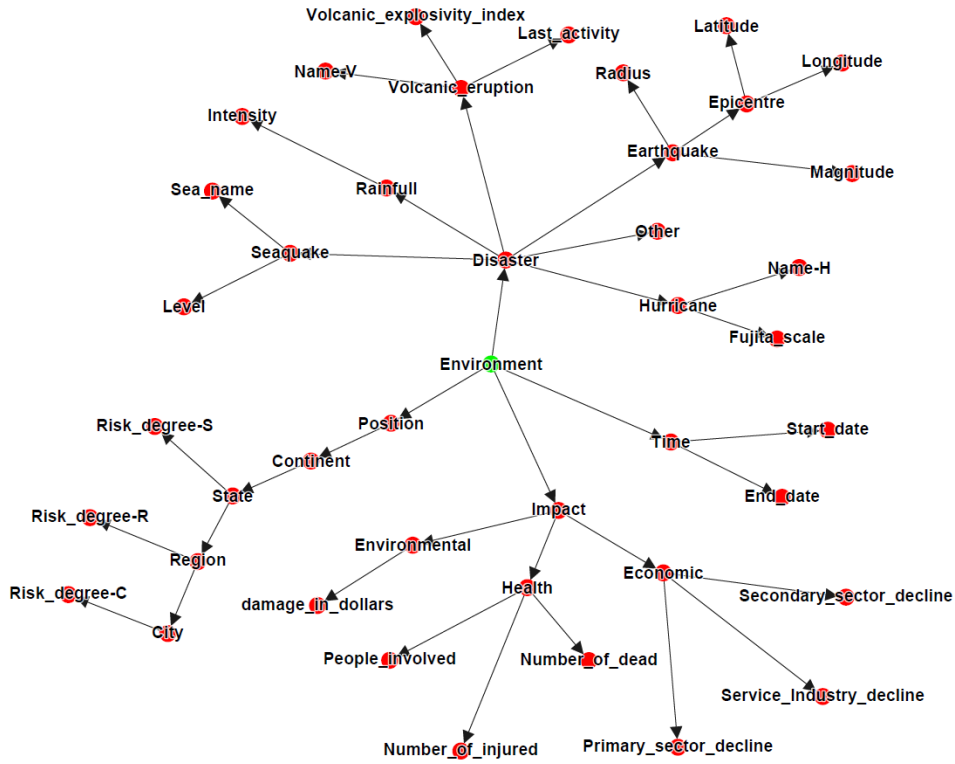


Fig. 7.4. The network corresponding to the source *Environment disasters*

pattern belong to the same network, the pattern is obtained simply by computing the shortest path from `Energy` to `Population` in the network of Figure 7.3. Actually, in

Keywords
$absorptivecapacity, action, adjustments, adopted, agencies, agricultural,$ $air, appraisal, areas, Bangladesh, Burton, capita, catastrophe, choice, coast,$ $alcomprehensive, coping, cost, crops, damage, deaths, developingcountries,$ $relief, drought, earthquake, economic, effects, effort, emergency, environment,$ $environmental, estimated, evacuation, experience, extremeeventsfarmers, federa,$ $Figure, flood, floodplain, forecasting, frequency, global, globalwarming, groups,$ $hazardevent, hazardresearch, human, hurricane, impact, income, increase, individual,$ $industrial, Kates, LabourBrigade, land, LiuLing, loss, magnitude, maize, major,$ $measures, ment, million, naturaldisasters, naturalevents, naturalhazards,$ $hazard, Nicaragua, occur, organization, pattern, people's communepercent, percent,$ $population, possible, potential, prevent, protection, range, reduce, regions,$ $risk, River, scale, scientific, social, society, SriLanka, storm, studies,$ $threshold, tion, tornado, TristandaCunha, tropical, cyclone, TropicalStormAgnes,$ $tsunami, UnitedKingdom, urban, vulnerable, warning, systems, zone, Managua,$ $air, plant, disaster, airpollution, natural, Tanzania, TropicalStormAgnes.$

Table 7.1. Keywords of the source *Environment risks*

Keywords
$acid, activatedsludge, activity, aerosol, airpollution, airquality, air,$ $anaerobicdigestion, approach, aquatic, areas, Assesment, atmosphere, biofuels,$ $carbon, catalyst, cause, chemical, chlorine, climatechange, combustion,$ $concentrations, contaminated, cycle, CycleAssessment, deposition, diesel,$ $dose, drinkingwater, ecosystem, effects, effluent, emissions, energy,$ $EnvironmentAgency, European, EuropeanCommission, EuropeanUnion,$ $exposure, Figure, fuel, gases, global, human, hydrocarbons, impacts, important,$ $industrial, landfill, legislation, levels, London, major, materials, measures,$ $models, monitoring, nanoparticles, nitrate, nitrogen, nitrogendioxide, nuisance,$ $operation, organic, oxidation, oxygen, ozone, particles, PBDEs, PCBs, pesticides,$ $plant, potential, radiation, radiativeforcing, radioactive, range, reaction,$ $recycling, reduce, regulation, regulatory, release, response, result, risk, sewage,$ $significant, sludge, soil, sources, species, standards, stratosphere, studies,$ $substances, sulfurdioxide, surface, temperature, toxic, transport, treatment,$ $typically, urban, vehicles, wastemanagement, ambient, biological, compounds,$ $Directive, engine, example, increase, metals, petrol, reactor, eutrophication.$

Table 7.2. Keywords of the source *Air pollution*

this case, we have only one possible path, shown in Figure 7.5. This path consists of 4 i-arcs, no c-arcs and no pw-arcs.

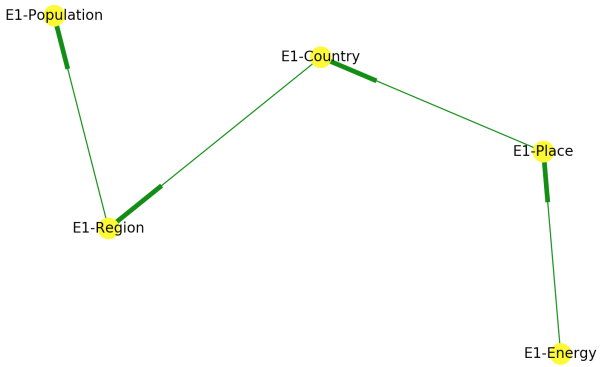


Fig. 7.5. Complex knowledge pattern from the node `Energy` to the node `Population` of the source *Energy*

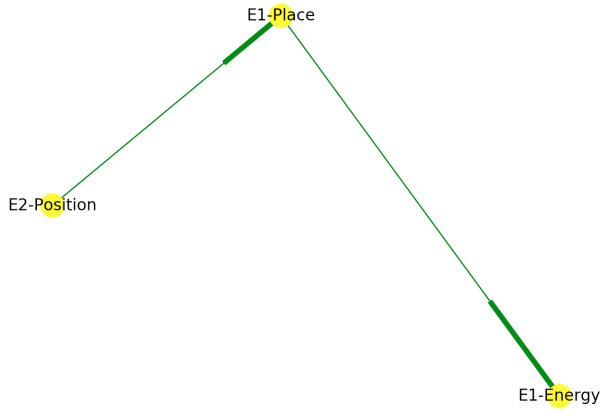


Fig. 7.6. Complex knowledge pattern from the node `Position` of the source *Environment disasters* to the node `Energy` of the source *Energy*

The second case study we are considering involves as "seed objects" `Position`, belonging to *Environment disasters*, and `Energy`, belonging to *Energy*. In this case, it is evident the necessity of passing through at least one c-arc because the two objects belong to different sources. One of the synonyms of the object `Position` is the object `Place`, belonging to the source *Energy*. As a consequence, it is possible to define at least one path, starting from `Position`, passing through `Place` and reaching `Energy`. This path is shown in Figure 7.6 and consists of 1 i-arc, 1 c-arc and no pw-arc. An alternative path would involve the nodes `Position` and `Continent` of *Environment disasters* and the nodes `Country`, `Place` and `Energy` of *Energy*. However, this path would consist of 3 i-arcs, 1 c-arc and no pw-arc and, clearly, it is not the shortest path. As a consequence, in this case, our approach returns the path shown in Figure 7.6 as the complex knowledge pattern from `Position` to `Energy`.

The third case study we are considering involves, as "seed objects", `Fujita_scale` of *Environment disasters* and `Risk` of *Environment risks*. In this case, synonymies are not sufficient because there is no synonymy involving `Fujita_scale`. However,

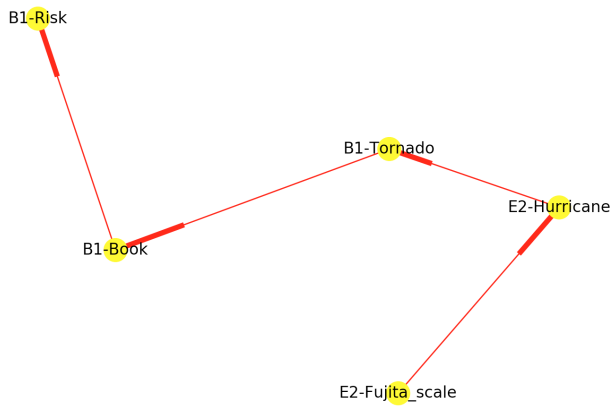


Fig. 7.7. Complex knowledge pattern from the node `Fujita_scale` of the source *Environment disasters* to the node `Risk` of the source *Environment risks*

the "whole" node of `Fujita_scale` is `Hurricane` and there is a synonymy between `Hurricane` and `Tornado`. As a consequence, it is possible to define at least one path starting from `Fujita_scale`, passing through `Hurricane` and `Tornado` and ending to `Risk`. This path is shown in Figure 7.7. It consists of 1 i-arc, 1 c-arc and 1 pw-arc. This is also the shortest path from `Fujita_scale` to `Risk` and, therefore, the complex knowledge pattern between these two nodes.

The fourth and last case study is the most complex one because it involves more alternative synonymies that could be selected, with the consequent need to determine the best one. The "seed objects" are `Risk_degree` of *Environment disasters* and `Emergency` of *Environment risks*. `Risk_degree` presents two synonymies in the dictionary; the former involves the object `Risk` of *Environment risks*; the latter regards the object `Risk` of *Air pollution*. As a consequence, at least two extremely different paths could exits. However, the path involving the node `Risk` of *Environment risks* can reach the target source through only 1 c-arc. The other one would need at least another c-arc to reach the target source. In particular, it should use the synonymy between `Risk` of *Air pollution* and `Hazard` of *Environment risks*. In Figure 7.8, we report both these paths. The former involves the nodes `Risk_degree`, `Risk`, `Book` and `Emergency` and consists of 2 i-arcs and 1 c-arc. The latter involves the nodes `Risk_degree`, `Risk`, `Hazard`, `Book` and `Emergency` and consists of 2 i-arcs and 2 c-arcs. Clearly, the shortest path is the former, which is selected as the complex knowledge pattern between the two seed nodes.

7.7 Discussion

This section is devoted to present a critical discussion of several aspects concerning our approach. It consists of four subsections. In the first, we present a comparison between

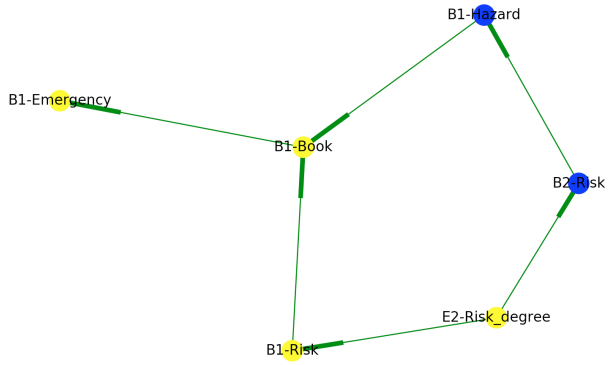


Fig. 7.8. Complex knowledge pattern from the node `Risk_degree` of the source *Environment disasters* to the node `Risk` of the source *Environment risks*

our approach and the related ones. In the second, we evaluate the performance of our technique for structuring unstructured data. In the third, we evaluate the performance of our overall approach. Finally, in the fourth, we measure its efficiency for large datasets. To carry out the experiments described in this section, we used a server equipped with an Intel I7 Dual Core 5500U processor and 16 GB of RAM with the Ubuntu 16.04.3 operating system. Clearly, especially for the last experiments, the capabilities of this server were limited. However, as we will see below, we were mostly interested to compare theoretical worst case response times with the real ones. Clearly, in real contexts, whenever necessary, much more powerful servers could be used to reduce the response time.

7.7.1 Comparison between our approach and the related ones

In Section 7.2, we have seen that we can distinguish four main families of approaches that are most related to ours. Specifically, these approaches aim at extracting: *(i)* keyword patterns; *(ii)* knowledge from structured sources; *(iii)* knowledge from heterogeneous sources; *(iv)* knowledge patterns through network analysis-based techniques.

As for the first family, the corresponding approaches share with ours the objective, i.e. the extraction of some form of knowledge. However, the knowledge derived by them consists simply in keyword patterns. Furthermore, the techniques they leverage to carry out this task are different from ours, especially if we consider the sub-family operating on RDF data. A higher similarity can be found with the other sub-family, i.e., the one operating on graph databases [116].

As for the second family, the corresponding approaches present some analogies, but also some differences, with ours. In particular, both of them exploit metadata to perform the knowledge extraction task. However, the approaches of this second family operate only on structured sources, whereas our approach manages sources with disparate formats.

The approaches of the third family extract knowledge from heterogeneous (both structured and semi-structured) sources. For instance, the approach of [274] aims at querying heterogeneous data in fuzzy XML documents by using a tree-pattern based algorithm. This approach has several differences with respect to ours. In fact, it focuses mainly on querying, whereas our approach considers the extraction of knowledge patterns. Furthermore, it operates on XML documents, whereas our approach operates on sources with different formats. Interestingly, the approach of [274] leverages a tree pattern-based algorithm, whereas ours exploits a graph pattern-based one. Another approach of this family is the one described in [402], which is based on Similarity Join. This approach and ours are similar in that both of them have a step in which a similarity detection task is performed. However, the approach of [402] needs a support knowledge hierarchy, whereas our approach exploits one or more thesauruses. Furthermore, the data sources considered by the approach of [402] are just collections of objects (e.g., documents) and not a variegate data lake, which is the reference data structure for our approach.

The fourth family comprises network-based models and algorithms that exploit network analysis to extract knowledge patterns. One of these approaches is described in [393]. It operates in the context of Music Information Retrieval, which is actually quite far from the scenarios of interest to our approach. However, both this approach and ours share the usage of network to represent available data and of network analysis to extract the desired knowledge. The approach of [393] focuses on non-traditional data sources, and, in this fact, is similar to ours. However, the source typology considered by it has a very specific nature, whereas the ones handled by our approach are numerous and are the most common ones encountered in the reality. Another approach belonging to the last family is the one described in [281]. This approach and ours present some analogies in that both of them use network analysis to extract knowledge of interest. However, the approach of [281] operates on only one kind of databases (e.g., relational ones) and focuses on a very specific topic, i.e., patent and applicant analysis. By contrast, our approach considers heterogeneous data formats and can operate on sources concerning different topics.

Other two approaches of this family that we have mentioned in Section 7.2 are the ones presented in [290] and [227]. [290] proposes a network-based formalism for representing available knowledge. In this formalism, nodes indicate concepts and arcs denote relationships between concepts. This representation coincides with the one adopted by our approach. The main difference between them consists in the fact that the approach of [290] operates only on information organized in structured databases. This fact contributes to perform data investigation and formalization very easily but, on the other hand, it prevents from managing semi-structured and unstructured data.

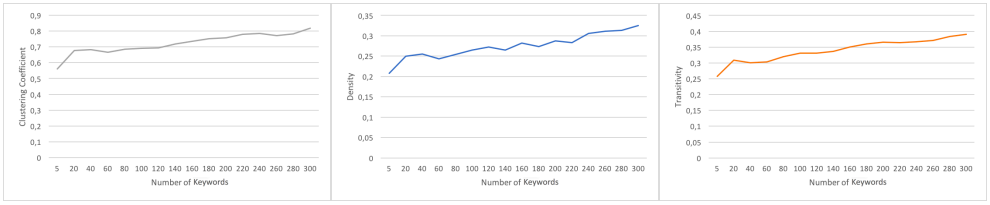


Fig. 7.9. Average clustering coefficient, density and transitivity of the network returned by our approach against the number of available keywords of the corresponding source

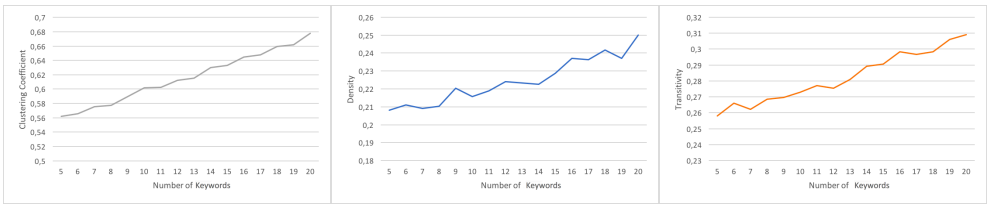


Fig. 7.10. A zoom of the graphs of Figure 7.9 referred to the case in which the number of keywords ranges between 5 and 20

The approach of [227] aims at performing keyword search in a graph to facilitate the identification of sub-graphs. This approach and ours are similar in that both of them are network-based. However, they also present several differences. Indeed, the algorithm underlying the approach of [227] is centered on cliques, whereas the one underlying our approach is based on paths. Furthermore, the approach of [227] focuses on keyword search, and the consequent subgraph identification, whereas ours aims at detecting knowledge patterns.

7.7.2 Evaluation of our approach to structure unstructured data

One way to evaluate the performance of our approach to structuring unstructured sources consists of determining how much it is able to connect the concepts corresponding to the flat keywords commonly used to characterize unstructured sources. Given a network-based model, like ours, a logical way to quantify this feature is based on the exploitation of some measures tipically adopted in network analysis to quantify the structuring level of a network. These measures are: *(i)* average clustering coefficient, *(ii)* density, and *(iii)* transitivity. All of them range in the real interval $[0, 1]$; the higher their value, the more structured the corresponding network.

We computed the values of these measures against the number of keywords representing unstructured sources. Obtained results are reported in Figure 7.9, whereas in Figure 7.10 we propose a "zoom" referred to the case in which the number of keywords ranges between 5 and 20.

From the analysis of these figures we can observe that our approach is really capable of structuring unstructured sources provided that the number of keywords

representing each source is higher than a reasonable minimum value (i.e., 5). As long as the number of provided keywords increases, the values of all our structuring capability indicators increase, even if this increase is very slow.

In our opinion, the growth slowness, far from being a problem, is an indicator of correctness. Indeed, we must consider that we are trying to assign a structure to an originally unstructured source. Our approach can provide a certain structuring level but it cannot (and it must not) upset the original nature of the source, which is unstructured.

All these reasonings allow us to say that our approach to structuring unstructured sources presents a very satisfying behavior.

7.7.3 Performance of our overall approach

In Section 7.5.2, we have seen that the computational complexity of the extraction of complex knowledge patterns is $O(|A| \cdot log|N| \cdot |N_{max}|)$. We have also seen that this complexity can be judged very satisfactory, if we consider the problem to solve.

However, in real data lakes, the number of involved sources is high and so, in principle, $|N|$ (and $|A|$, which is $O(|N|^2)$ could be very high. Nevertheless, in real situations, the number of relationships among attributes and elements is very small and, consequently, the corresponding networks are very sparse. As a consequence, $|A|$ should be very low, if compared with $|N|^2$, and, therefore, we were confident that, in real cases, the performance of our approach should be very good.

To verify this hypothesis we measured the response time of our approach when the number of involved nodes to examine increases; in particular, we measured the response time obtained by considering the theoretical computational complexity and the real response time. Obtained results are reported in Figure 7.11, whereas in Figure 7.12 we propose a "zoom" for those cases that in Figure 7.11 appeared superimposed on the axis of abscissas. In these graphs, in the computation of the theoretical response time, we considered several values of graph density.

From the analysis of these figures, it clearly emerges that, in real cases, the response time of our approach is much smaller than the one determined by the worst case time complexity, even when the network density is low or very low. This fact leads our approach to work very well also in presence of large data lakes, provided that the corresponding networks are sparse or very sparse, which is the general condition that is found in practice. As a consequence, we can conclude that our hypothesis was true and, therefore, that our approach shows a good performance in real scenarios.

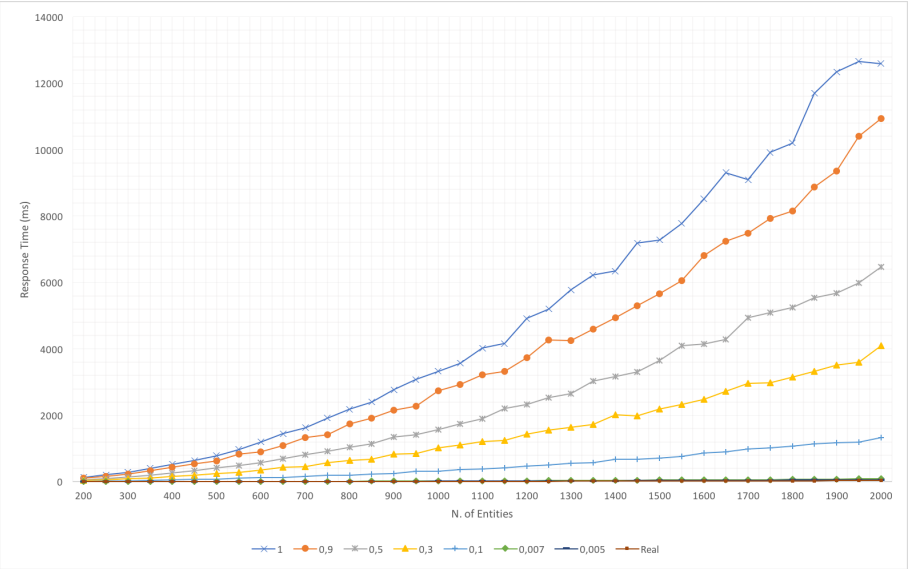


Fig. 7.11. Real and theoretical response time against data lake dimension and density

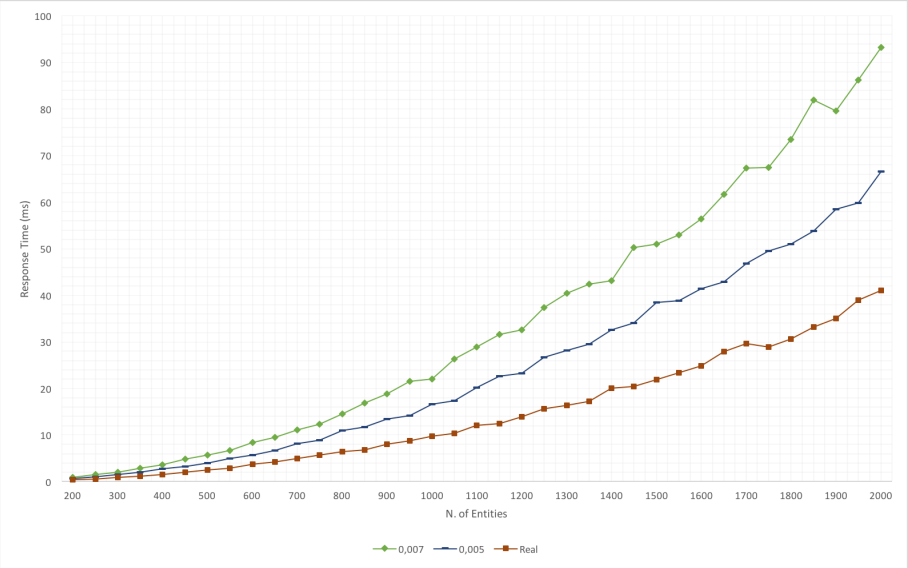


Fig. 7.12. Real and theoretical response time against data lake dimension and density (zoom of Figure 7.11)

7.7.4 Efficiency of our overall approach for large data sets

In Section 7.5.2, we have seen that, from a theoretical point of view, in order to determine the computational complexity of our approach, we must consider two main scenarios, namely:

1. the detected path involves nodes of only one source, in which case the theoretical computational complexity is $O(|A| \cdot log(|N|)$;
2. the detected path involves nodes of more sources, in which case the theoretical computational complexity is $O(|A| \cdot log|N|) \cdot O(max(|N_{max}|, |DL|))$.

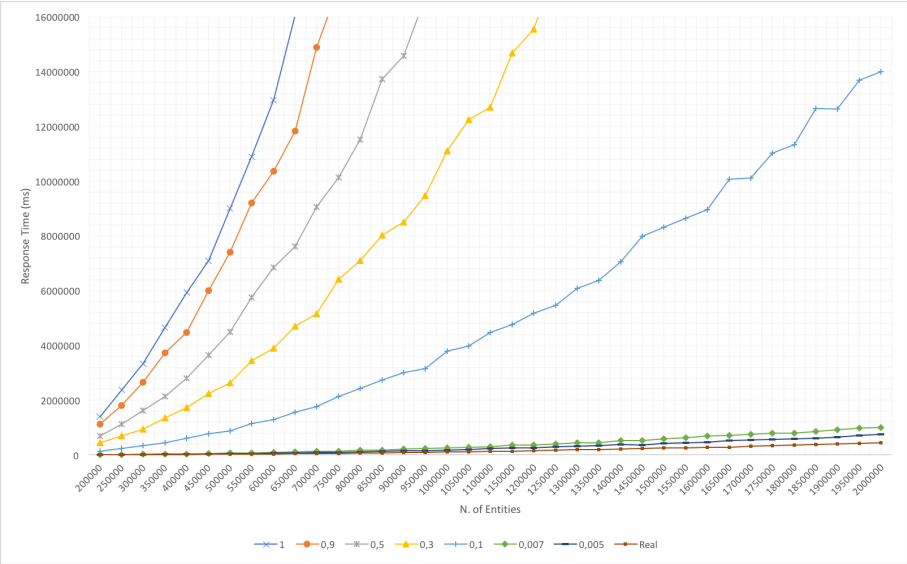


Fig. 7.13. Real and theoretical response time against dimension and density for large data lakes (Scenario 1)

Now, in presence of large data lakes, both $|N_{max}|$ and $|DL|$ are much smaller than $|N|$; as a consequence, from a theoretical point of view, the two cases could be referred to a single one. However, since we aim at measuring the efficiency of our approach in the reality (and not only from a theoretical viewpoint), we prefer to keep the two cases separate and to verify if this hypothesis is also confirmed in practice.

To carry out this experiment, we decided to repeat the tasks already performed in the previous one, but with a data lake having a number of nodes that is three orders of magnitude higher than the maximum one considered in the previous experiment. This number of nodes is clearly much higher than the ones we can currently meet in real situations. However, we preferred to put our approach under stress to see if, even in these extreme cases, it shows an acceptable behavior. Also in this case, we computed the response time of our approach against the number of nodes of the data lake and compared the response time obtained by considering the theoretical computational complexity against the real response time.

Obtained results are reported in Figure 7.13, for the Scenario 1 mentioned above, and in Figure 7.15, for the Scenario 2 considered previously. A "zoom" of these figures, limited to those cases that appeared superimposed on the axis of abscissas, are reported in Figures 7.14 and 7.16, respectively. From the analysis of these figures we can observe that, in presence of very large data sets, the theoretical response time of our approach would make it not applicable for high values of density. Instead, our approach shows an acceptable response time for low values of density.

Actually, we have already seen that, in real cases, data lake density is very low. This is also witnessed by the trend of the real response time shown in Figures 7.13 - 7.16,

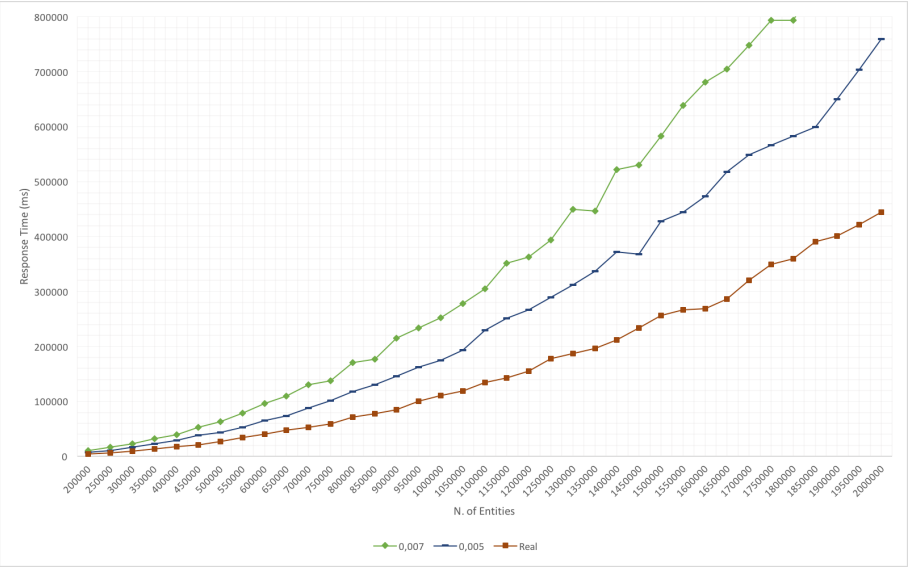


Fig. 7.14. Real and theoretical response time against dimension and density for large data lakes (zoom of Figure 7.13)

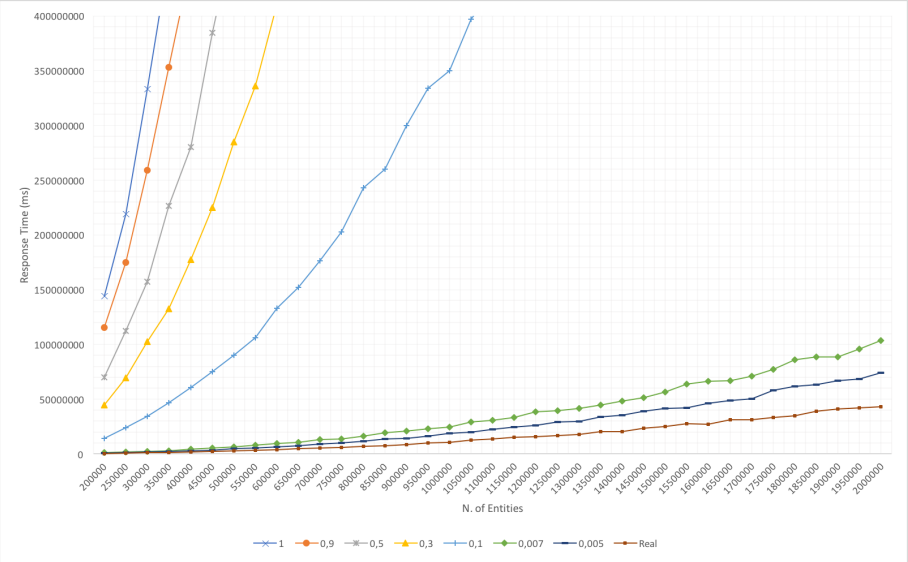


Fig. 7.15. Real and theoretical response time against dimension and density for large data lakes (Scenario 2)

which is even better than the response time derived from the theoretical computational complexity obtained with a small density (i.e., 0.005). Interestingly, the trends of the real response time for Scenarios 1 and 2 are actually the same. The only difference regards the corresponding values that, in case of Scenario 2, are about two orders of magnitude higher than the ones shown in Scenario 1.

All the results described in this section, coupled with the fact that we stressed our approach in extreme cases generally not found in the current reality, lead us to

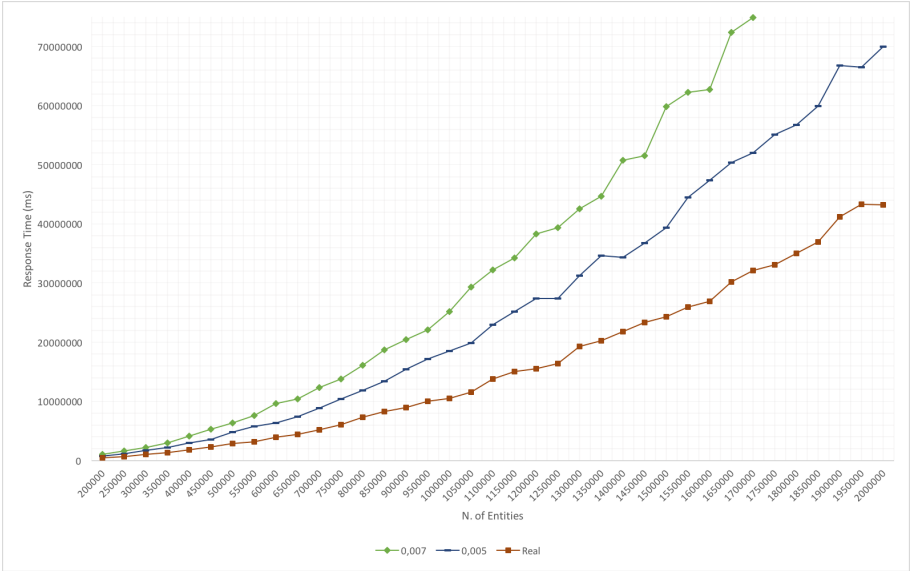


Fig. 7.16. Real and theoretical response time against data lake dimension and density for large data lakes (zoom of Figure 7.15)

conclude that our approach presents a very good efficiency, which makes it well suited also for large datasets.

Part III

Internet of Things

In this part, we apply our network-based model and the associated social network-based approach to IoT and we propose a new social IoT paradigm, called MIoT. A Multiple-IoT can be seen as a set of things connected to each other by relationships of any kind and, at the same time, as a set of related IoTs, one for each kind of relationship. This part is organized as follows: in Chapter 8, we present an approach to extract knowledge from heterogeneous sensor data streams. In Chapter 9, we present the MIoT paradigm. Finally, in Chapter 10, we introduce the concept of topic-guided virtual IoTs in a MIoT.

8

Extracting knowledge from heterogeneous sensor data streams

8.1 Introduction

In the last few years, research on Wireless Sensor Networks (WSNs) has been ignited by important advances in various technological areas, such as wireless communications, digital electronics and micro-electro-mechanical systems. These improvements allowed for an easy development of low-power and low-cost multi-functional sensors and networks thereof. Sensor networks usually include a large number of nodes, each of which may sense several measures. Cooperation among nodes is usually sought for in such networks. Sensor nodes are usually positioned either inside or very close to observed events, and the main objective is to provide users with a better understanding of the environment in which sensors are deployed, thus giving the opportunity to acquire new information and intelligence. While the management of sensor networks and the development of robust data acquisition layers received much attention in the literature, one big open challenge in this research area is anomaly detection [74, 75]. Anomalies can be generated by either malfunctioning sensors or changes in the monitored environment. In most cases, being able to distinguish between the two scenarios is a challenging task. Most of the past approaches for anomaly detection focused on the analysis of data produced by each single device [477]. The most notable approaches in this setting can be grouped in four categories, namely: *(i)* rule-based detection [204], *(ii)* statistical techniques [263], *(iii)* graph-based techniques [331], and *(iv)* data mining and computational intelligence-based techniques [471]. Instead, network-based approaches for anomaly detection in WSNs received less attention [114, 216, 113, 165]. In fact, in spite of a strict complementarity and correlation between network analysis and WSNs, only in the latest years, researchers have begun to apply network analysis-based techniques to WSNs. However, they have only proposed the application of classical network analysis parameters to this context. Indeed, most of the proposed approaches employ centrality measures [350], which allow the detection of anomalies of only one node at a time.

In this chapter, we aim at introducing new solutions for the analysis of heterogeneous sensors organized as a network. In particular, our techniques will be based on the evaluation of the connectivity of the whole WSN and its subnetworks (instead of on node centrality), and are mainly focused on potential anomalies involving more sensors located therein. They adopt a metric capable of uniformly handling measures provided by heterogeneous sensors, as well as a dashboard of network analysis parameters. This way, they allow the detection of anomalies involving more (heterogeneous) sensors, and the evaluation of the impact of these anomalies on the whole sensor network and its subnetworks. The plan of this chapter is as follows. In Section 8.2, we introduce our model used to represent WSNs and our anomaly detection approach. In Section 8.3, we present some preliminary results on tests carried out on a sensor network, along with some discussions.

8.2 Methods

8.2.1 Network construction

Let $\mathcal{W}$ be a WSN. Without loss of generality, assume that the corresponding sensors can be partitioned along two orthogonal dimensions[1]. In the scenario considered here, these dimensions are location and physical quantities to evaluate (in particular, we consider $p = 3$ physical quantities, i.e., temperature, lightness and humidity). Assume that the WSN covers l locations (in particular, we consider $l = 3$ locations, named A, B and C in the following) and that one location contains n devices, each measuring p physical quantities. As a consequence, the overall number of sensors is $s = pln$.

A network $\mathcal{N} = \langle V, E \rangle$ can be associated with $\mathcal{W}$. Here, V is the set of the nodes of $\mathcal{N}$. Each node $v_i \in V$ corresponds to a sensor and has associated a label $\langle l_i, p_i \rangle$, where l_i represents its location and p_i denotes the physical quantity it measures. E is the set of the edges of $\mathcal{N}$. Each edge e_{ij} connects the nodes v_i and v_j. It can be represented as $e_{ij} = (v_i, v_j, w_{ij})$. Here, w_{ij} is a measure of "distance" between v_i and v_j. It is an indicator of the non-correlation level of the sensors associated with v_i and v_j. Actually, each parameter representing this feature could be adopted in our model. In the experiments presented in this chapter we adopted Multi-Parameterized Edit Distance (MPED) [91] for its capability of measuring the non-correlation level of sensors regarding heterogeneous physical quantities, characterized by different units of measure and possible data shifts.

$\mathcal{N}$ can be partitioned along one or both dimensions. We indicate by $\mathcal{N}_p = \langle V_p, E_p \rangle$ the subnets obtained by taking only the nodes that correspond to the sensors mea-

[1] Actually, the number of dimensions could be greater than two, without requiring any change of the approach.

suring the physical quantity p. Here, $p \in \{l, t, h\}$ can denote lightness, temperature and humidity, respectively. Analogously, we indicate by $\mathcal{N}_q = \langle V_q, E_q \rangle$ the subnets obtained by taking only the nodes that correspond to the sensors operating at the location q. Here, $q \in \{A, B, C\}$. Finally, we denote by $\mathcal{N}_{pq} = \langle V_{pq}, E_{pq} \rangle$ the subnet obtained by considering only the nodes corresponding to the sensors that measure the physical quantity p and operate in the location q, along with the edges linking them.

8.2.2 Network parameters

As pointed out in the Introduction, we use several parameters to construct our dashboard supporting the extraction of knowledge about environment changes. The first four parameters are derived from classical network theory; the fifth is derived from a particular centrality measure proposed in [243]; the last is introduced by us. In this section, we present an overview of these parameters. In the following, we define all of them on a reference network $\mathcal{N} = \langle V, E \rangle$. The first parameter is the *Characteristic Path Length*, also known as the *Average Shortest Path Length*. It is defined as the average length of the shortest paths connecting all possible pairs of network nodes. More formally, let $l(v_i, v_j)$ be the length of the shortest path between v_i and v_j. The Characteristic Path Length $\mathcal{L}_\mathcal{N}$ of $\mathcal{N}$ is defined as: $\mathcal{L}_\mathcal{N} = \frac{1}{|V|(|V|-1)} \sum_{v_i \in V} \sum_{v_j \in V, v_j \neq v_i} l(v_i, v_j)$. The second parameter is the *Average Node Connectivity*. Given two nodes v_i and v_j, their connectivity $c(v_i, v_j)$ represents the minimum number of edges that need to be removed to disconnect them. The Average Node Connectivity $\mathcal{C}_\mathcal{N}$ is defined as: $\mathcal{C}_\mathcal{N} = \frac{1}{|V|(|V|-1)} \sum_{v_i \in V} \sum_{v_j \in V, v_j \neq v_i} c(v_i, v_j)$. The third parameter is the *Average Number of Simple Paths*. Given two nodes v_i and v_j, we indicate by $p(v_i, v_j)$ the number of simple paths (i.e., paths with no repeated nodes) between them. Then, we define the Average Number of Simple Paths $\mathcal{P}_\mathcal{N}$ as: $\mathcal{P}_\mathcal{N} = \frac{1}{|V|(|V|-1)} \sum_{v_i \in V} \sum_{v_j \in V, v_j \neq v_i} p(v_i, v_j)$. The fourth parameter is the *Average Clustering Coefficient*. In order to define it, we must preliminarily introduce the neighborhood $nbh(v_i)$ of a node v_i as follows: $nbh(v_i) = \{v_j | e_{ij} \in E\}$. Then, we define the Clustering Coefficient of a node v_i as: $s(v_i) = \frac{2 \cdot |\{e_{jk} | v_j, v_k \in nbh(v_i), e_{jk} \in E\}|}{|nbh(v_i)| \cdot (|nbh(v_i)|-1)}$. Finally, we define the Average Clustering Coefficient as: $\mathcal{S}_\mathcal{N} = \frac{1}{|V|} \sum_{v_i \in V} s(v_i)$. The fifth parameter is the *Average Closeness Vitality*. Given a node v_i, the closeness vitality $t(v_i)$ represents the increase in the sum of distances between all the pairs of nodes of $\mathcal{N}$, when v_i is excluded from $\mathcal{N}$ [243]. The Average Closeness Vitality $\mathcal{T}_\mathcal{N}$ is defined as: $\mathcal{T}_\mathcal{N} = \frac{1}{|V|} \sum_{v_i \in V} t(v_i)$. The sixth parameter (i.e., the one introduced by us) is the *Connection Coefficient*. It starts from the observation that, in network analysis, one of the most powerful tools for investigating the connection level of a network is the concept of clique. As a consequence, it is reasonable to adopt this concept to evaluate the cohesion of a network. This coefficient takes the following considerations into

account: *(i)* both the dimension and the number of cliques are important as connectivity indicators; *(ii)* the concept of clique is intrinsically exponential; in other words, a clique of dimension $n+1$ is exponentially more complex than a clique of dimension n.

In order to define the Connection Coefficient it is necessary to introduce a support network $\mathcal{N}^{\pi} = \langle V, E^{\pi} \rangle$, obtained by removing from $\mathcal{N}$ the edges with an "excessive" weight; observe that the nodes of $\mathcal{N}^{\pi}$ are the same as the nodes of $\mathcal{N}$. To formally define E^{π}, we employ the distribution of the weights of the edges of $\mathcal{N}$. Specifically, let max_E (resp., min_E) be the maximum (resp., minimum) weight of an edge of E. It is possible to define a parameter $step_E = \frac{max_E - min_E}{10}$, which represents the length of a "step" of the interval between min_E and max_E. We can define $d^k(E)$, $0 \leq k \leq 9$, as the number of the edges of E whose weights belong to the interval between $min_E + k \cdot step_E$ and $min_E + (k+1) \cdot step_E$. All these intervals are closed on the left and open on the right, except for the last one that is closed both on the left and on the right. E^{π} can be defined as: $E^{\pi} = \{e_{ij} \in E | e_{ij} \in \bigcup_{k \leq th_{max}} d^k(E)\}$. We have experimentally set $th_{max} = 6$. We are now able to define the Connection Coefficient $\mathcal{Q}_{\mathcal{N}}$ of $\mathcal{N}$. In particular, let C be the set of the cliques of $\mathcal{N}^{\pi}$; let C_k be the set of cliques of dimension k of $\mathcal{N}^{\pi}$; finally, let $|C_k|$ be the cardinality (i.e., the number of cliques) of C_k. Then, $\mathcal{Q}_{\mathcal{N}}$ is defined as: $\mathcal{Q}_{\mathcal{N}} = \sum_{k=1}^{|V|} |C_k| \cdot 2^k$.

8.2.3 Approach to knowledge extraction

The idea underlying our approach is that, if some changes occur on sensor data streams, then some variations can be observed in some or all the dashboard parameters, when measured on the whole network, and/or on some of its subnetworks, depending on the number, the kind and the location of involved sensors. Our approach consists of a training phase and a testing phase. To carry out them, we employed available data (see Section 8.3.1) and, according to the holdout technique, we partitioned these data in such a way as to use 2/3 of them for the training phase and 1/3 of them for the testing phase. As for the training phase, we considered the following situations: *(1)* all sensors behaved correctly; *(2)* two sensors in location A and two sensors in location B were perturbed, in such a way as to decrease humidity; *(3)* two sensors in location B and two sensors in location C were perturbed, in such a way as to decrease lightness; *(4)* two sensors in location A and two sensors in location C were perturbed, in such a way as to increase lightness. Obtained results, along with the corresponding discussion, are presented in Section 8.3. After the training phase, we started the testing phase. In this case, we considered the following situations: *(1)* all sensors behaved correctly; *(2)* two sensors in location B and two sensors in location C were perturbed, in such a way as to decrease humidity; *(3)* two sensors in location

A and two sensors in location C were perturbed, in such a way as to decrease lightness; *(4)* two sensors in location A and two sensors in location B were perturbed, in such a way as to increase lightness. Obtained results, along with the corresponding discussion, are presented in Section 8.3. Here, we simply point out that our approach behaved very well and was capable of correctly identifying all perturbations.

Finally, we applied our approach to the following situations: *(1)* one sensor in the location A and one sensor in the location B were perturbed, in such a way as to decrease humidity; *(2)* one sensor in the locations A and C was perturbed, in such a way as to increase lightness, and one sensor in the locations B and C was perturbed, in such a way as to decrease the same physical quantity; *(3)* three sensors in the location A and one sensor in the location B were perturbed, in such a way as to decrease humidity; *(4)* one sensor in the location A was perturbed, in such a way as to increase humidity; *(5)* one sensor in the location B was perturbed, in such a way as to increase lightness. Obtained results, along with the corresponding discussion, are presented in Section 8.3. Here, we anticipate that our approach showed its suitability to detect almost all perturbations.

8.3 Results

8.3.1 Testbed

To collect data for the experiments introduced in Section 8.2.3, we built a WSN by following specific guidelines. In particular, we organized devices in a multi-hop Wireless Sensor Area Network (WSAN) and managed them through the Building Management Framework (BMF) [159]. This is a framework for domain-specific networks, which offers an efficient and flexible management of WSANs deployed in indoor areas by allowing users to take advantage of sensing/actuation intelligent techniques and fast prototyping of WSAN applications. BMF enabled the use of heterogeneous WSANs through a base station, which acted both as data collector and network configurator. Communication between base station and devices was carried out by means of the BMF Communication Protocol, an application level protocol built on top of multi-hop network protocols [261, 172]. We composed the WSAN using MICAz sensor devices, providing 128 kB for program storage, 512 kB for data storage, and 4 kB of RAM. Devices were powered mainly by means of external power. They were configured to communicate with the base station, sending data every minute. To test our approach, we synthetically injected several anomalies at pre-determined time slots. In particular, to increase lightness, we employed artificial sources of lightness with controlled intensity, whereas to reduce lightness, we applied artificial lightness filters. Finally, humidity was controlled by chemicals. Our network consisted of 9 devices labeled by

increasing numbers. Each device included 3 sensors, which retrieved values for humidity, lightness and temperature. Devices 1, 2 and 3 have been positioned in location A, devices 4, 5 and 6 operated in location B, devices 7, 8 and 9 were situated in location C. A, B and C were three different rooms on the same floor of a building. Finally, we collected data for 24 days without perturbations and other 36 days with several perturbations, as described in Section 8.2.3.

8.3.2 Obtained results and Discussion

In this section, we report the results obtained by performing all the experiments mentioned in Section 8.2.3. Preliminarily, we observe that the definition of the six coefficients forming our dashboard suggests that a decrease of the connection level of a network or a subnetwork leads to: *(i)* an increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$; *(ii)* a decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$. The purpose of the training phase was to find the optimal values of some thresholds underlying our approach (for instance, the value of th_{max} in the definition of Connection Coefficient - see Section 8.2.2) and to have a first idea of its behavior. In Table 8.1, we report all the results regarding the training phase after the optimal values of thresholds were set. In particular, this table consists of four sub-tables, each corresponding to one of the four situations mentioned in Section 8.2.3. For each situation, we report the values of the six parameters of the dashboard for the overall network and the subnetworks $\mathcal{N}_t$, $\mathcal{N}_l$, $\mathcal{N}_h$, $\mathcal{N}_A$, $\mathcal{N}_B$ and $\mathcal{N}_C$ (see Section 8.2.1). In this table, Situation 1 represents the correct one. In Situation 2, we observe: *(i)* a very high increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$, along with a very high decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$ for the network $\mathcal{N}_h$; *(ii)* a high increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$, along with a high decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$ for the networks $\mathcal{N}_A$ and $\mathcal{N}_B$; *(iii)* a moderate increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$, along with a moderate decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$ for the overall network. In Situation 3 (resp., 4), we observe: *(i)* a very high increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$, along with a very high decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$ for the network $\mathcal{N}_l$; *(ii)* a high increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$, along with a high decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$ for the networks $\mathcal{N}_B$ and $\mathcal{N}_C$ (resp., $\mathcal{N}_A$ and $\mathcal{N}_C$); *(iii)* a moderate increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$, along with a moderate decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$ for the overall network. These results confirm that our approach is really capable of capturing the perturbations in wireless sensor networks or subnetworks caused by sensor anomalies (and, indirectly, it is able to evaluate the network and subnetwork resilience to sensor anomalies). The only weakness revealed by this first test is that, in its current version, our approach is not able to tell us if these perturbations are caused by an increase or a decrease of the corresponding physical quantity.

The purpose of the testing phase was to verify both the setting of the threshold values and the corresponding results detected during the training phase. In Table 8.2,

$Network$	$\mathcal{L}_{\mathcal{N}}$	$\mathcal{C}_{\mathcal{N}}$	$\mathcal{P}_{\mathcal{N}}$	$\mathcal{T}_{\mathcal{N}}$	$\mathcal{Q}_{\mathcal{N}}$	$\mathcal{S}_{\mathcal{N}}$
$Overall$	1.1054	22.4387	6508290	64.2548	1163264	0.8944
$\mathcal{N}_t$	1.0322	7.1056	14232	15.1429	592	0.8413
$\mathcal{N}_l$	1.0451	7.1111	13200	16.6667	592	0.8595
$\mathcal{N}_h$	1.0278	7.5833	16758	16.9143	512	0.9722
$\mathcal{N}_A$	1.1944	5.6944	8012	23.7241	224	0.8339
$\mathcal{N}_B$	1.1667	5.9444	9274	22.4000	256	0.8582
$\mathcal{N}_C$	1.1944	6.0556	7896	23.7241	288	0.7794
$Overall$	1.1795	20.0684	4652472	74.7500	227328	0.8239
$\mathcal{N}_t$	1.1189	6.4444	10376	21.1613	384	0.8212
$\mathcal{N}_l$	1.1011	6.5833	11816	20.0000	320	0.7905
$\mathcal{N}_h$	1.4167	3.9444	2268	38.0952	96	0.5270
$\mathcal{N}_A$	1.3611	4.5000	3208	34.0870	120	0.5582
$\mathcal{N}_B$	1.3456	4.7778	4572	32.0800	144	0.5858
$\mathcal{N}_C$	1.1833	6.0444	7828	26.9091	248	0.7832
$Overall$	1.2194	19.1937	3790486	81.2263	99840	0.7796
$\mathcal{N}_t$	1.2556	5.8778	9924	20.8824	412	0.7392
$\mathcal{N}_l$	1.5000	4.1111	6102	26.3704	192	0.6000
$\mathcal{N}_h$	1.0556	7.2778	14924	17.8824	512	0.9392
$\mathcal{N}_A$	1.2111	5.4000	7990	23.0000	200	0.8571
$\mathcal{N}_B$	1.3222	4.5278	5990	29.1429	108	0.5630
$\mathcal{N}_C$	1.3333	4.7778	3824	32.0000	120	0.5407
$Overall$	1.2394	18.1937	3480632	80.2263	97650	0.7823
$\mathcal{N}_t$	1.2356	5.6648	9633	21.2435	408	0.7491
$\mathcal{N}_l$	1.5200	3.9345	6260	27.3221	192	0.5800
$\mathcal{N}_h$	1.0776	6.9318	13924	17.7623	512	0.9154
$\mathcal{N}_A$	1.3782	4.4987	5843	28.2322	108	0.661
$\mathcal{N}_B$	1.1911	5.1000	7232	23.0000	206	0.8200
$\mathcal{N}_C$	1.3433	4.6578	3126	31.6850	120	0.5207

Table 8.1. Results obtained by our approach during the training phase

we report all the results regarding this phase. Observe that the situations considered during this phase are the same as the ones examined during the training phase; however, we modified the subnetworks (among A, B and C) involved in each perturbation in such a way as to prevent overfitting. Obtained results confirm that the selection of the threshold values performed during the training phase was correct. They also confirm all the observations about the features of our approach, which we drew at the end of the training phase.

After the testing phase confirmed the suitability of our approach, we applied it to new situations not considered during the previous phases. These situations are described in detail in Section 8.2.3. In Table 8.3, we report the corresponding results. From their analysis we can draw very interesting observations. In particular, in Situation 1, we obtain the same trend as the one seen in Situation 2 of the training phase. However, the perturbation degree is more reduced. This is correct because, for locations A and B, we perturbed one sensor, instead of two. In Situation 2, we observe: *(i)* a very high increase of $\mathcal{L}_{\mathcal{N}}$ and $\mathcal{T}_{\mathcal{N}}$, along with a very high decrease of $\mathcal{C}_{\mathcal{N}}$, $\mathcal{P}_{\mathcal{N}}$, $\mathcal{S}_{\mathcal{N}}$ and $\mathcal{Q}_{\mathcal{N}}$ for the network $\mathcal{N}_l$; these increases and decreases are comparable with the ones observed in Situation 3 of the training phase; *(ii)* a moderate (resp., high,

$Network$	$\mathcal{L}_\mathcal{N}$	$\mathcal{C}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{T}_\mathcal{N}$	$\mathcal{Q}_\mathcal{N}$	$\mathcal{S}_\mathcal{N}$
$Overall$	1.1135	20.4387	7120293	65.3746	1163264	0.9144
$\mathcal{N}_t$	1.0411	6.5306	13939	15.1529	592	0.8712
$\mathcal{N}_l$	1.0361	6.2480	13737	17.1227	592	0.8891
$\mathcal{N}_h$	1.0235	7.3311	16123	16.8242	512	0.8920
$\mathcal{N}_A$	1.1826	5.4129	7910	22.7241	228	0.8451
$\mathcal{N}_B$	1.1700	5.8331	8992	21.4000	256	0.8112
$\mathcal{N}_C$	1.1929	6.2410	7786	23.7241	288	0.8042
$Overall$	1.1896	20.1224	4993459	72.63	294629	0.8484
$\mathcal{N}_t$	1.1289	6.2468	11001	22.1982	320	0.8391
$\mathcal{N}_l$	1.2133	6.6631	10829	21.0782	384	0.8081
$\mathcal{N}_h$	1.5177	3.8104	3124	37.1719	112	0.5328
$\mathcal{N}_A$	1.1922	6.2324	7128	27.8801	208	0.7312
$\mathcal{N}_B$	1.3232	4.9188	4492	31.9500	128	0.5558
$\mathcal{N}_C$	1.3511	4.4780	3198	33.0870	118	0.5182
$Overall$	1.2766	20.2308	4290486	81.3094	97744	0.7824
$\mathcal{N}_t$	1.3111	5.5833	9850	20.0000	258	0.7825
$\mathcal{N}_l$	1.4389	4.0833	3438	25.9750	96	0.6412
$\mathcal{N}_h$	1.0242	7.3611	13978	18.4421	384	0.9825
$\mathcal{N}_A$	1.3056	4.5278	4762	30.1515	108	0.5713
$\mathcal{N}_B$	1.1896	5.5278	7288	22.1429	216	0.8462
$\mathcal{N}_C$	1.2825	4.9444	3594	32.9143	96	0.5356
$Overall$	1.2251	17.9876	3990563	82.2263	97650	0.7769
$\mathcal{N}_t$	1.2944	5.8326	9112	22.7241	408	0.7839
$\mathcal{N}_l$	1.4678	4.6161	6383	26.3352	112	0.5455
$\mathcal{N}_h$	1.1111	6.5833	13816	17.6686	384	0.9005
$\mathcal{N}_A$	1.4001	4.7144	6152	27.8652	96	0.6148
$\mathcal{N}_B$	1.3675	4.3056	3886	30.9850	88	0.5198
$\mathcal{N}_C$	1.1887	6.2421	7341	22.7692	256	0.8825

Table 8.2. Results obtained by our approach during the testing phase

very high) increase of $\mathcal{L}_\mathcal{N}$ and $\mathcal{T}_\mathcal{N}$, along with a moderate (resp., high) decrease of $\mathcal{C}_\mathcal{N}$, $\mathcal{P}_\mathcal{N}$, $\mathcal{S}_\mathcal{N}$ and $\mathcal{Q}_\mathcal{N}$, for the networks $\mathcal{N}_A$ and $\mathcal{N}_B$ (resp., $\mathcal{N}_C$, $\mathcal{N}_l$); *(iii)* a moderate increase of $\mathcal{L}_\mathcal{N}$ and $\mathcal{T}_\mathcal{N}$, along with a moderate decrease of $\mathcal{C}_\mathcal{N}$, $\mathcal{P}_\mathcal{N}$, $\mathcal{S}_\mathcal{N}$ and $\mathcal{Q}_\mathcal{N}$, for the overall network. Observe that, since our approach considers perturbations, but it currently does not distinguish between increases and decreases, even if, in the network $\mathcal{N}_l$, there are opposite perturbations in two lightness sensors, their consequences are not nullified by our approach, but, on the contrary, are "combined" by it. In our opinion, this is a correct behavior of our approach. In Situation 3, we observe: *(i)* an increase (resp., decrease) of $\mathcal{L}_\mathcal{N}$ and $\mathcal{T}_\mathcal{N}$ (resp., $\mathcal{C}_\mathcal{N}$, $\mathcal{P}_\mathcal{N}$, $\mathcal{S}_\mathcal{N}$ and $\mathcal{Q}_\mathcal{N}$), comparable with the one of Situation 2 of the training phase for both the overall network and the network $\mathcal{N}_h$; *(ii)* a significant (resp., moderate) increase of $\mathcal{L}_\mathcal{N}$ and $\mathcal{T}_\mathcal{N}$, along with a significant (resp., moderate) decrease of $\mathcal{C}_\mathcal{N}$, $\mathcal{P}_\mathcal{N}$, $\mathcal{S}_\mathcal{N}$ and $\mathcal{Q}_\mathcal{N}$ for the network $\mathcal{N}_A$ (resp., $\mathcal{N}_B$). In Situation 4 (resp., 5), we observe: *(i)* a very moderate increase of $\mathcal{L}_\mathcal{N}$ and $\mathcal{T}_\mathcal{N}$, along with a very moderate decrease of $\mathcal{C}_\mathcal{N}$, $\mathcal{P}_\mathcal{N}$, $\mathcal{S}_\mathcal{N}$ and $\mathcal{Q}_\mathcal{N}$ for the overall network and for the networks $\mathcal{N}_h$ and $\mathcal{N}_A$ (resp., $\mathcal{N}_l$ and $\mathcal{N}_B$). This reveals a second weakness of our approach, which shows a difficulty to find a single anomaly. Indeed, in this case, it found a slight change in the dashboard parameters for both

$Network$	$\mathcal{L}_{\mathcal{N}}$	$\mathcal{C}_{\mathcal{N}}$	$\mathcal{P}_{\mathcal{N}}$	$\mathcal{T}_{\mathcal{N}}$	$\mathcal{Q}_{\mathcal{N}}$	$\mathcal{S}_{\mathcal{N}}$
$Overall$	1.1435	21.5534	5580928	70.0000	114264	0.8534
$\mathcal{N}_t$	1.0712	6.3159	11432	18.2221	384	0.8613
$\mathcal{N}_l$	1.0572	6.4354	11202	18.6667	384	0.8564
$\mathcal{N}_h$	1.2578	4.5673	4564	22.2124	144	0.8123
$\mathcal{N}_A$	1.2235	5.1843	6006	28.3673	200	0.7034
$\mathcal{N}_B$	1.2351	5.4992	8842	27.4332	224	0.6982
$\mathcal{N}_C$	1.1833	5.3556	7828	24.7347	248	0.7792
$Overall$	1.2199	19.3747	3948573	80.3252	97650	0.7856
$\mathcal{N}_t$	1.2456	5.6658	8562	21.9383	388	0.7467
$\mathcal{N}_l$	1.6100	3.5039	5987	28.2392	192	0.5971
$\mathcal{N}_h$	1.0877	6.4837	12527	17.3877	512	0.8672
$\mathcal{N}_A$	1.2292	4.5948	7873	27.223	228	0.6823
$\mathcal{N}_B$	1.2334	5.1229	7367	26.2391	228	0.6891
$\mathcal{N}_C$	1.2921	4.6578	3834	32.2320	120	0.5012
$Overall$	1.1235	21.9987	3977283	74.5673	231872	0.8223
$\mathcal{N}_t$	1.1312	6.2989	12345	21.3939	512	0.8323
$\mathcal{N}_l$	1.1433	6.5643	12234	20.3332	512	0.8340
$\mathcal{N}_h$	1.4872	3.9440	3542	38.9412	120	0.7795
$\mathcal{N}_A$	1.8342	2.2338	1987	35.1843	96	0.4032
$\mathcal{N}_B$	1.2151	4.4738	6932	25.6230	224	0.5820
$\mathcal{N}_C$	1.1933	6.0872	8239	23.3235	284	0.7780
$Overall$	1.1228	21.3789	6184736	67.3233	131872	0.8534
$\mathcal{N}_t$	1.0613	6.4599	12341	17.3939	592	0.8613
$\mathcal{N}_l$	1.0732	6.8865	12854	16.3452	592	0.8564
$\mathcal{N}_h$	1.1640	5.6534	9532	20.9482	288	0.8123
$\mathcal{N}_A$	1.2132	5.1928	6987	26.1212	288	0.7034
$\mathcal{N}_B$	1.1951	5.4738	9928	24.7210	320	0.6982
$\mathcal{N}_C$	1.19445	5.5872	8239	23.3235	320	0.7792
$Overall$	1.1289	21.8729	6857326	67.3252	131662	0.8556
$\mathcal{N}_t$	1.0782	6.7654	12662	17.2352	592	0.8467
$\mathcal{N}_l$	1.1728	5.9987	5987	20.4568	288	0.8023
$\mathcal{N}_h$	1.0654	6.2356	12277	16.4555	592	0.8553
$\mathcal{N}_A$	1.1892	5.6457	9854	25.3356	320	0.7061
$\mathcal{N}_B$	1.2234	5.0101	5346	26.4564	288	0.7072
$\mathcal{N}_C$	1.1921	5.5482	8899	23.2845	284	0.7843

Table 8.3. Results obtained by our approach during the examination of some situations of interest

the whole network and the involved subnetworks. This is mainly due to the purpose of our approach, which does not aim at performing anomaly detection in one sensor (actually, a long list of approaches carrying out this task - e.g., [204, 263, 331, 471] - already exists) but, instead, it aims at detecting the consequences, on the whole network and its subnetworks, of anomalies involving more (heterogeneous) sensors installed in different locations. In fact, in this case, the interaction of these anomalies in the network could be extremely variegate and could depend on the number, the kind and the location of perturbed sensors, so that their detection, along with the detection of their effects, becomes extremely difficult and justifies the employment of quite time-expensive approaches like ours. As for this issue, the results described in this section allow us to conclude that our approach reaches the objectives for which it was designed.

9

Multiple IoTs

9.1 Introduction

We already pointed out in Section 1 that the idea underlying SIoT is extremely interesting and, as a matter of fact, has received, and is still receiving, a lot of attention in the literature. However, we think that, in the next future, the number of relationships that might connect things could be much higher than five, and relationships could be much more variegate than the ones currently considered by SIoT. As a consequence, we think that a new paradigm, taking into account this fact, is in order.

In [82, 335], we introduced the concept of Social Internetworking System (SIS, for short) as a system comprising an undefined number of users, social networks and resources. The SIS paradigm was thought to extend the Single Social Network paradigm by taking into account that: *(i)* a user can join many social networks, *(ii)* these joins can often vary over time, and *(iii)* the presence of users joining more social networks can favor the cooperation of users, who do not join the same social networks. We think that the key concepts of SIS can also be applied to things (instead of to users) and to relationships between things and, in this chapter, we propose the MIoT (Multiple Internets of Things) paradigm. The core of the SIS paradigm is modeling users and their relationships as a unique big network and, at the same time, as a set of related social networks connected to each other thanks to those users joining more than one social network. In this chapter, we propose to extend the ideas underlying the concept of SIS to IoT. The MIoT paradigm arises as a result of this objective.

Roughly speaking, a MIoT can be seen as a set of things connected to each other by relationships of any kind and, at the same time, as a set of related IoTs, one for each kind of relationship. Actually, a more precise definition of MIoT would require the introduction of the concept of instance of a thing in an IoT. According to this concept, the instance of a thing in an IoT represents a virtual view of that thing in the IoT. Having this in mind, a MIoT can be seen as a set of related IoTs, one for each kind of relationship into consideration. The nodes of each IoT represent the

instances of the things participating to it. As a consequence, a thing can have several instances, one for each IoT to which it participates. As will be clear in the following, the existence of more instances for one thing plays a key role in the MIoT paradigm because it allows the definition of the cross relationships among the different IoTs of the MIoT.

Differently from SIoT, in the MIoT paradigm, the number of relationships is not defined a priori. In a MIoT, there is a node for each thing; furthermore, there is an edge between two nodes if the corresponding things are linked by a relationship. If more kinds of relationship exist between two things, then more edges exist between the corresponding nodes, one for each kind of relationship. All the nodes linked by a given kind of relationship, together with the corresponding edges, form an IoT of the MIoT.

Observe that, under this MIoT definition, SIoT can be seen as a specific case of MIoT in which the number of the possible kinds of relationship is limited to 5 and these kinds are pre-defined. IoTs are interconnected thanks to those nodes corresponding to things involved in more than one kind of relationship. We call *cross nodes* (*c-nodes*, for short) these nodes and *inner nodes* (*i-nodes*, for short) all the other ones. Then, a c-node connects at least two IoTs of the MIoT and plays a key role to favor the cooperation among i-nodes belonging to different IoTs. As a consequence, differently from SIoT, the nodes of a MIoT are not all equal: c-nodes will presumably play a more important role than i-nodes for supporting the activities in a MIoT.

Note that the MIoT paradigm can be seen as an attempt to address an open issue evidenced in [40] about some improvements that should be made on the SIoT paradigm. Among these improvements, two very relevant ones evidenced in this chapter are the following:

- defining inter-objects relationships; this issue requires a correct representation of a smart object and the definition of both methods and tools to crawl and discover other (possibly heterogeneous) objects with which interactions can be established;
- modeling the new social networks thus obtained, characterizing them and defining new algorithms to perform their analysis.

The MIoT paradigm already mentioned, and the crawling strategy, which we present below, taken together, can represent an answer to these exigencies of improvement.

From a more applicative point of view, having some IoTs that can "communicate" through c-nodes can lead to some beneficial synergies. For instance, assume that an environment-related IoT can communicate with a home-related IoT through a cross node. Assume that the former IoT evidences an abnormal presence of dioxin in a place

located some kilometers away from the home (for instance, owing to a fire of a plastic deposit). Assume, also, that this IoT is evidencing that the wind direction is pushing the dioxin towards the home. The home-related IoT could be "informed" through a cross node about this fact and could close all windows before the arrival of the dioxin.

Once a MIoT has been defined, it is possible to apply Social Network Analysis-based techniques on it to extract powerful knowledge concerning its things, their relationships, the IoTs formed by them, etc. However, in order to perform knowledge extraction, especially when the number of the things to investigate is huge, an important pre-requisite is having a good approach to crawl the underlying graph. Crawling is also extremely useful in a second family of applications, based on the exploration of the "neighborhood" (i.e., things and relationships) of a given thing (think, for instance, of the case in which a new thing is added to the Internet of Things and wants to create relationships with other things). There are also a lot of further possible applications of crawling, already known in the literature [338, 436], and that can be extended to the Internet of Things.

In the literature, several crawling strategies for single social networks have been proposed. Among them, the most representative ones are: *(i)* Breadth-First Search (BFS, for short) [470], which moves in breadth by exploring the neighborhood of each node; *(ii)* Random Walk (RW, for short) [283], which moves in random directions; *(iii)* Metropolis-Hastings Random Walk (MH, for short) [426, 247, 376], which moves in random directions, disfavoring high-degree nodes. These strategies were largely investigated for single networks, and their pros and cons have been highlighted in [171, 249].

However, we have seen that, in a MIoT, there exist two different kinds of node, and none of the previous strategies considers this fact, as they were developed for crawling a single network. We argue that a new strategy, capable of distinguishing c-nodes from i-nodes and of performing a right tradeoff between breadth, depth and randomicity, is in order. Therefore, a second objective of this chapter is addressing this issue. In fact, we propose a new crawling strategy, called *Cross Node Driven Search* (CDS, for short). CDS is centered on c-nodes; in fact, it allows users to privilege the visit of c-nodes over the one of i-nodes, if necessary, and to tune how much c-nodes should be privileged over i-nodes.

To prove the correctness of CDS, we tested it against the three main classic strategies mentioned above. In carrying out this task, we defined, and, then, used different metrics aimed to evaluate the quality of each crawler under consideration. The results of these experiments confirm our assumption about the inadequacy of the classic crawling strategies for a MIoT and, by contrast, the suitability of the new CDS strategy in this context.

This chapter is organized as follows: in Section 9.2, we illustrate related literature. In Section 9.3, we present the MIoT paradigm. In Section 9.4, we describe the CDS crawler and illustrate the experimental campaign, which we performed to test it. In Section 9.5 we propose a comparison between our model and approach and other, more or less conventional, ones.

9.2 Related Literature

Several years have passed since the IoT paradigm was introduced [38, 41, 314, 363]. During this period, the term "Internet of Things – IoT" has been associated with a huge variety of concepts, technologies and solutions. For instance, in the last few years, new technologies, such as Big Data [86] and Social Networking, have been applied to IoT and have changed, and are currently changing, the very definition of this term. What IoT will become in the future depends on the evolution of these technologies [438].

The current research on IoT focuses on the capability of connecting every object to the Internet. This way of thinking IoT led to the Web of Things (hereafter, WoT) paradigm [184, 183, 215] and to the application of Social Networking to the IoT domain [40]. In the next future, these technologies will be combined with other ones, such as Information Centric Networks [425, 479, 480, 32, 372, 33, 361] and Cloud [131, 433, 229]. As a matter of fact, the strengths of these last ones are exactly the features necessary to overcome the weaknesses of the current IoT concept [467]. Some examples of this combination can be already found in the literature [150, 180, 449, 448].

Significant efforts have been made to apply the Social Networking ideas to the IoT domain. Actually, the implementation of reliable IoTs [39] passes through the definition of a complex architecture capable of managing services. In this research direction, the authors of [362] propose CASCOM, a model devoted to simplify the interaction between consumers and data in an IoT context. It is also necessary that this complex architecture enables a complete connectivity among things [248], guarantees quick reactions to frequent state variations and, finally, ensures a good scalability.

Furthermore, as IoT is based on the Internet, it must address the same security issues characterizing this network [222]. Therefore, the development of new architectures capable of fulfilling security and privacy requirements is in order [483].

The first attempts to apply Social Networking to the IoT domain can be found in [182, 333, 245, 205]. In these papers, the authors propose to use human social network relationships to share services provided by a set of things.

An important step forward is performed in [39], where the SIoT paradigm is introduced. Here, the authors propose an approach to creating relationships among things, without requiring the owner intervention. Thanks to this idea, things can autonomously crawl the network to find services and resources of their interest provided by other things. In [42], the same authors clearly highlight what are the main strengths of SIoT. Specifically: *(i)* the SIoT structure can be dynamically modified to ensure network navigability and to find new things; *(ii)* scalability is guaranteed, like in human social networks; *(iii)* a level of trustworthiness between things can be established; *(iv)* the past social network approaches can be redefined to solve problems typical of the IoT context [342].

Today, the connection level of humans and things is continuously increasing, so that it appears reasonable to start to investigate the "network of networks" scenario, thus passing from Social Networking to Social Internetworking. One of the most interesting attempts in this direction is Social Internetworking System (hereafter, SIS); it regards the connection of several human networks to form a network of human networks [82, 335]. The strength of SIS resides in the fact that this structure is capable of interconnecting users joining different social networks. In this new scenario, concepts and tools of Social Network Analysis can be adapted to evaluate the main features concerning the interactions between users belonging to the same network or to different networks. This new paradigm aims at guaranteing a tradeoff between the autonomy of each network of the SIS and the possibility of increasing power, efficiency and effectiveness, obtained through the interaction of the networks of the SIS. To the best of our knowledge, no architecture similar to SIS has been proposed for networks of things yet.

In [40], the authors point out that there are still several open issues that must be investigated in the SIoT paradigm. In particular, making things capable of establishing heterogeneous social relationships requires specific investigations and new approaches. Among them, the most relevant ones for our context are: *(i) Defining inter-objects relationships*. This task requires a correct digital representation of a smart object and the definition of a methodological and technological solution capable of crawling and discovering other (possibly heterogeneous) objects, with which interactions can be established. *(ii) Modeling the new social graphs thus obtained*, in such a way as to characterize them and to define new algorithms for performing their analysis.

Crawling represents a key issue for the implementation of the IoT paradigm. The necessity of addressing this issue is mentioned in many papers (e.g., [40, 286, 404, 168, 133], to cite a few). In spite of this high demand, just few papers addressing this problem can be found in the past literature on IoTs. Most of the approaches proposed in these papers focus on the creation of search engines conceived to operate on IoT

[286, 289] or, more often, on the Web of Things [436, 104]. In [436], an accurate survey on this last research area is presented.

In [154], the authors propose a geo-based crawler for IoT aiming at minimizing inter-site communication costs. Every site uses its own crawler that is provided with some predefined rules for fetching and parsing the Web. In [133], a framework to automatize the search, and the next classification, of services belonging to a digital health ecosystem, is proposed. This framework exploits both a focused web crawler, which explores the network, and a social classification system. In [269], the authors propose an approach aimed at improving the existing web crawlers, when they operate on IoT, and to catch up the fingerprints of the IoT nodes. This approach is based on an incremental crawler, which periodically classifies nodes in such a way as to ensure the highest classification accuracy for the most important ones.

In [266], the crawling problem is approached from a different perspective. Indeed, one of the main problems in a network of things is battery consumption. To avoid it, in most cases, sensors perform a working-sleeping duty cycle. The authors of [266] model the crawling problem as a scheduling one and define a sleep-aware schedule method called EasiCrow. This method is well suited to crawl sensors with an asynchronous sleeping cycle. In [403], the authors, starting from the assumption that things are becoming the major producers and consumers of data, propose a system to extract data from different sources. Once data has been acquired, this system provides suitable interfaces allowing both humans and machines to share and dynamically search the services of their interest.

9.3 The MIoT paradigm

We define a MIoT $\mathcal{M}$ as a set of m Internets of Things (see Figure 9.1 for a schematic representation of it)[1]. Formally speaking:

$$\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$$

where $\mathcal{I}_k$ is an IoT.

Let o_j be an object of $\mathcal{M}$. We assume that, if o_j belongs to $\mathcal{I}_k$, it has an instance ι_{j_k}, representing it in $\mathcal{I}_k$. As pointed out in the Introduction, in this chapter, the instance ι_{j_k} indicates a virtual view (or, better, a virtual agent) representing o_j in $\mathcal{I}_k$.

[1] In this chapter, the term "IoT" is intended according to the new trends that characterize this research field [40]. These trends suggest that, with the explosion of the number of available things, it is not realistic to talk about a unique Internet of Things. By contrast, it is more appropriate to consider several IoTs, each consisting of a (social) network of things.

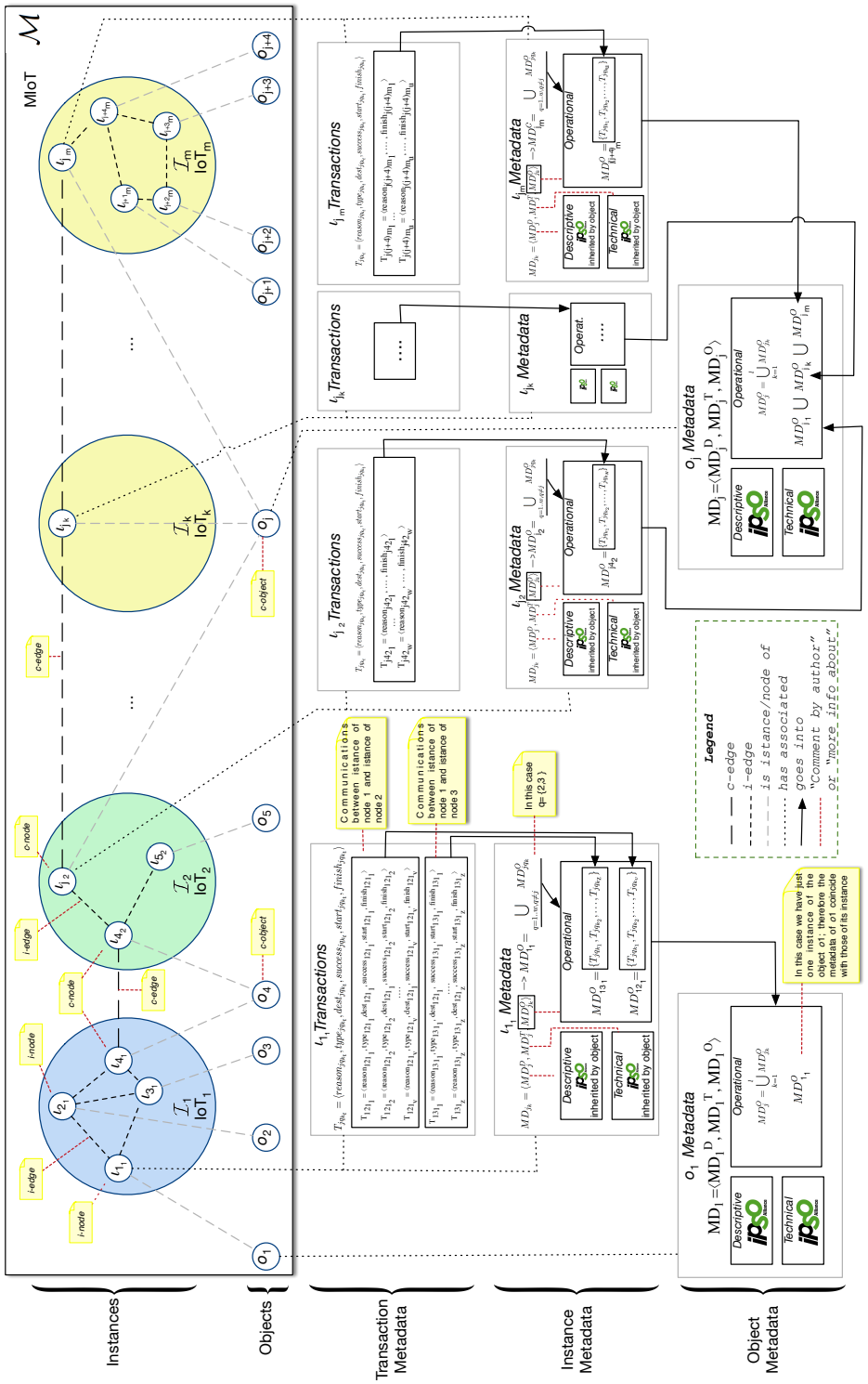


Fig. 9.1. Schematic representation of the proposed MIoT structure

For instance, it provides all the other instances of $\mathcal{I}_k$, as well as the users interacting with $\mathcal{I}_k$, with all necessary information about o_j. Interestingly, this information is represented according to the format and the conventions adopted in $\mathcal{I}_k$.

In $\mathcal{M}$, *a set* MD_j *of metadata* are associated with an object o_j. We define a rich set of metadata of an object, because these play a key role in favoring the interoperability of IoTs and of their objects, which is the main objective of a MIoT. As a consequence, MD_j consists of three different subsets:

$$MD_j = \langle MD_j^D, MD_j^T, MD_j^O \rangle$$

Here:

- MD_j^D represents the set of *descriptive metadata*. It denotes the type of o_j. For representing and handling descriptive metadata, a proper taxonomy, such as the one defined by the IPSO Alliance [5], can be adopted.
- MD_j^T represents the set of *technical metadata*. It must be compliant with the object type. In other words, there is a different set of metadata for each object type of the taxonomy. Also in this case, the IPSO Alliance provides a well defined set of technical metadata for each object type. It is worth pointing out that, in principle, we could have allowed much richer descriptive and technical metadata. However, we did not make this choice because we preferred to relate our definition of metadata to an international IoT standard, such as the one defined by the IPSO Alliance. Furthermore, as will be clear in the following, our approach needs mainly operational metadata. As a consequence, making descriptive and technical metadata more complex would have added a useless level of complexity to our model.
- MD_j^O represents the set of *operational metadata*. It regards the behavior of o_j. The operational metadata of an object o_j is defined as the union of the sets of the operational metadata of its instances. Specifically, let $\iota_{j_1}, \iota_{j_2}, \ldots, \iota_{j_l}$, $l \leq m$, be the instances of o_j belonging to the IoTs of $\mathcal{M}$. Then:

  $$MD_j^O = \bigcup_{k=1}^{l} MD_{j_k}^O$$

  $MD_{j_k}^O$ is the set of the operational metadata of the instance ι_{j_k}. In order to understand the structure of $MD_{j_k}^O$, we first have to analyze the structure of $MD_{jq_k}^O$, i.e. the set of operational metadata between two instances ι_{j_k} and ι_{q_k}, of the objects o_j and o_q, in the IoT $\mathcal{I}_k$.

  Specifically, $MD_{jq_k}^O$ is given by the set of metadata associated with the transactions between ι_{j_k} and ι_{q_k}. In particular:

  $$MD_{jq_k}^O = \{T_{jq_{k_1}}, T_{jq_{k_2}}, \ldots, T_{jq_{k_v}}\}$$

where $T_{jq_{k_t}}$, $1 \leq t \leq v$, represents the metadata of the t-th transaction between ι_{j_k} and ι_{q_k}, assuming that v is the current number of transactions between the two instances.

$T_{jq_{k_t}}$ can be represented as follows:

$$T_{jq_{k_t}} = \langle reason_{jq_{k_t}}, type_{jq_{k_t}}, inst1_{jq_{k_t}}, inst2_{jq_{k_t}}, success_{jq_{k_t}}, start_{jq_{k_t}}, finish_{jq_{k_t}} \rangle$$

where:

- $reason_{jq_{k_t}}$ denotes the reason causing the transaction, chosen among a set of default values.
- $type_{jq_{k_t}}$ indicates the transaction type (e.g., unicast, multicast, and so forth).
- $inst1_{jq_{k_t}}$ and $inst2_{jq_{k_t}}$ denote the two instances involved in $T_{jq_{k_t}}$. Observe that a transaction between ι_{j_k} and ι_{q_k} could be part of a longer path whose source and/or target nodes could be different from ι_{j_k} and ι_{q_k}. In principle, the source and/or the target nodes of a transaction could belong to an IoT different from $\mathcal{I}_k$. In this last case, it is necessary to reach $\mathcal{I}_k$ from the source, and/or to reach the target from $\mathcal{I}_k$, through one or more cross nodes, if possible.
- $success_{jq_{k_t}}$ denotes if the transaction succeeded.
- $start_{jq_{k_t}}$ is the timestamp associated with the beginning of the transaction.
- $finish_{jq_{k_t}}$ is the timestamp associated with the end of the transaction (its value is NULL if $T_{jq_{k_t}}$ failed).

In our model, the direction of a transaction is not considered. Furthermore, the parameter v, i.e., the number of transactions for each pair of instances, varies when moving from a pair of instances to another.

Observe that we have made our model powerful enough to represent and handle all the transactions between two instances of each IoT. Having all these detailed historical data at disposal could help the analysis of the real "social" behavior of each object. Furthermore, these data could be exploited in many applications; think, for instance, of the computation of the trust and reputation of each object, the investigation of objects with similar or complementary behaviors, and so forth. On the other hand, maintaining a full history of transactions may be very expensive and useless in many real life applications; in some cases, suitable data summarizations could be enough. As a consequence, when passing from the abstract model definition to real life applications, the transaction representation could be removed, extended or restricted on the basis of a tradeoff between costs and benefits for the current application.

We are now able to define the set of the operational metadata MD^O_{jk} of an instance ι_{j_k} of $\mathcal{I}_k$. Specifically, let $\iota_{1_k}, \iota_{2_k}, \ldots, \iota_{w_k}$ be all the instances belonging to $\mathcal{I}_k$. Then:

$$MD^O_{jk} = \bigcup_{q=1..w, q \neq j} MD^O_{jq_k}$$

In other words, the set of the operational metadata of an instance ι_{j_k} is given by the union of the sets of the operational metadata of the transactions between ι_{j_k} and all the other instances of $\mathcal{I}_k$.

Given an instance ι_{j_k}, relative to an object o_j and an IoT $\mathcal{I}_k$, we define the metadata MD_{j_k} of ι_{j_k} as:

$$MD_{j_k} = \langle MD^D_j, MD^T_j, MD^O_{j_k} \rangle$$

In other words, the descriptive and the technical metadata of an instance ι_{j_k} coincide with the ones of the corresponding object o_j. Instead, the operational metadata of ι_{j_k} is a subset of the operational metadata of o_j that comprise only those ones regarding the transactions, which ι_{j_k} is involved in.

It is possible to associate a graph:

$$G_k = \langle N_k, A_k \rangle$$

with $\mathcal{I}_k$. Here, N_k indicates the set of the nodes of $\mathcal{I}_k$. There is a node n_{j_k} for each instance ι_{j_k} of an object o_j in $\mathcal{I}_k$. A_k denotes the set of the edges of $\mathcal{I}_k$. There is an edge $a_{jq_k} = (n_{j_k}, n_{q_k})$ if there exists a link between the instances ι_{j_k} and ι_{q_k} of the objects o_j and o_q in the IoT $\mathcal{I}_k$.

Also the overall MIoT $\mathcal{M}$ can be represented as a graph:

$$\mathcal{M} = \langle N, A \rangle$$

Here:

- $N = \bigcup_{k=1}^{m} N_k$;
- $A = A_I \cup A_C$, where:
  - $A_I = \bigcup_{k=1}^{m} A_k$;
  - $A_C = \{(n_{j_k}, n_{j_q}) | n_{j_k} \in N_k, n_{j_q} \in N_q, k \neq q\}$; observe that n_{j_k} and n_{j_q} are the nodes corresponding to the instances ι_{j_k} and ι_{j_q} of the object o_j in $\mathcal{I}_k$ and $\mathcal{I}_q$.

In other words, a MIoT $\mathcal{M}$ can be represented as a graph whose set of nodes is the union of the sets of nodes of the corresponding IoTs. The set A of the arcs of $\mathcal{M}$ consists of two subsets, A_I and A_C. A_I is the set of the inner arcs of $\mathcal{M}$ and is the union of the sets of the arcs of the corresponding IoTs. A_C is the set of the cross arcs of $\mathcal{M}$; there is a cross arc for each pair of instances of the same object in different IoTs. We call:

- *i-edge* an edge of $\mathcal{M}$ belonging to A_I;

- *c-edge* an edge of $\mathcal{M}$ belonging to A_C;
- *c-node* a node of $\mathcal{M}$ involved in at least one c-edge;
- *i-node* a node of $\mathcal{M}$ not involved in any c-edge;
- *c-object* an object having at least one pair of instances whose corresponding nodes are linked by a c-edge; clearly, any object with at least two different instances is a c-object.

It is worth pointing out that, as mentioned in the Introduction, there is a strict correlation between the MIoT paradigm and the concept of Social Internetworking System (hereafter, SIS) already presented in the literature [82]. In particular: *(i)* the concept of c-edges shares several features with the one of "me"-edge in a SIS; *(ii)* the concept of c-node is similar to the one of bridge in a SIS; *(iii)* a c-object corresponds to a user joining more social networks.

9.3.1 An example of a MIoT

Since the MIoT paradigm is new, in the Internet there is no known case study or real example about it yet. As a consequence, to provide the reader with an example, and, at the same time, to have a testbed for our experiments, we constructed a MIoT starting from some open data about things available on the Internet. In particular, we derived our data from *Thingful* [3]. This is a search engine for the Internet of Things, which allows us to search among a huge number of existing things, distributed all over the world. Thingful also provides some suitable APIs allowing the extraction of all the data we are looking for.

In order to construct our MIoT, we decided to work with 250 things whose data was derived from Thingful. Given the huge number of things available in Thingful, it could appear that the number of things composing our testbed is excessively limited. However, we observe that:

- This was the first attempt to construct a real MIoT and, then, it was extremely important for us to have a full control of it in order to verify if we were proceeding well. A full human control with a much higher number of nodes was not possible.
- We wanted to fully analyze the behavior, the strengths and the weaknesses of our crawler and to understand, step by step, its way of operating vs the ones of other crawlers. Again, a full human verification of these aspects was not possible with a larger testbed.
- As it will be clear in the following, our approach to obtaining the testbed is fully scalable. As a consequence, an interested researcher can apply it to construct a much larger testbed, if necessary.

We considered three dimensions of interest for our MIoT, namely:

a. *Category*: It specifies the application field which a given thing operates in. The categories we have chosen were five, namely *home*, *health*, *energy*, *transport*, and *environment*. Each category originated an IoT. Each thing was assigned to exactly one category.
b. *Coastal distance*: It specifies the coastal distance (i.e., the distance from any sea, lake or river) of each thing. The distance values we have set were:
   - *near*, for things distant less than 20 kilometres from the coast, for the categories *environment* and *energy*, and less than 5 kilometres, for the other three categories;
   - *mid*, for things whose minimum distance from the coast was between 20 and 105 kilometres, for the categories *environment* and *energy*, and between 5 and 25 kilometres, for the other three categories;
   - *far*, for things whose minimum distance from the coast was higher than 105 kilometres, for the categories *environment* and *energy*, and higher than 25 kilometres, for the other three categories.

   An IoT was created for each distance value. The different coastal distance values for *environment* and *energy*, on the one hand, and for the other three categories, on the other hand, have been determined after having analyzed the distribution of the involved categories of things against the coastal distance, in such a way as to produce a uniform distribution of each category of things in the three IoTs related to the coastal distance dimension.
c. *Altitude*: it specifies the altitude of the place where the thing is located. The altitude values we have defined were: *plain* (corresponding to an altitude less than 500 meters), *hill* (corresponding to an altitude between 500 and 1000 meters), and *mountain* (corresponding to an altitude higher than 1000 meters). An IoT was created for each altitude value.

As a consequence, our MIoT consists of 11 IoTs. We associated an object with each thing; therefore, we had 250 objects. In principle, for each object, we could have associated an instance for each dimension. However, in order to make our testbed closer to a generic MIoT, representing a real scenario, where it is not said that all the objects have exactly the same number of instances, we decided not to associate three instances with each object. Instead, we associated only one instance (distributed uniformly at random among the three dimensions, and based on the features of the things of the IoTs of a given dimension) to 200 of the 250 objects. Analogously, we associated two instances (distributed by following the same guidelines mentioned above) to 35 of the 250 objects. Finally, we associated three instances, one for each

IoT	*Number of instances*
a.home	22
a.health	22
a.energy	22
a.transport	22
a.environment	22
b.near	14
b.mid	38
b.far	53
c.plain	44
c.hill	50
c.mountain	6

Table 9.1. Number of instances present in the IoTs of our MIoT

possible dimension, to 15 of the 250 objects. At the end of this phase, we had 315 instances, distributed among the 11 IoTs of our MIoT as shown in Table 9.1.

To complete our MIoT and its network representation, we had to define a policy to create *i-edges*. In fact, it was clear that our MIoT should have had a node for each instance and a c-edge for each pair of instances referring to the same object. Therefore, the last decision regarded how to define i-edges. Given our scenario, it appeared reasonable to consider distances among things as the leading parameter for the creation of i-edges. To carry out this last task, we have preliminarily computed the distribution of the number of connected components possibly created from our instances against the maximum possible distance. Obtained results are reported in Figure 9.2. Based on this figure, in order to obtain a balanced number of connected components, we decided to connect two instances of the same IoT if the distance of the corresponding things was lesser than 1000 kilometres.

After this last choice, our MIoT was fully defined. In order to help the reader to mentally portray it, in Figure 9.3, we provide a graphical representation. The interested reader can find the corresponding dataset (in the `.csv` format) at the address `www.barbiana20.unirc.it/miot/datasets/miot1`. The password to type is "`za.12&;lq74:#`".

9.3.2 Why use the MIoT paradigm?

In the Introduction, we have specified that the MIoT paradigm goes in the direction suggested by some authors, who observe that it is no longer possible to think of a single global Internet of Things [40].

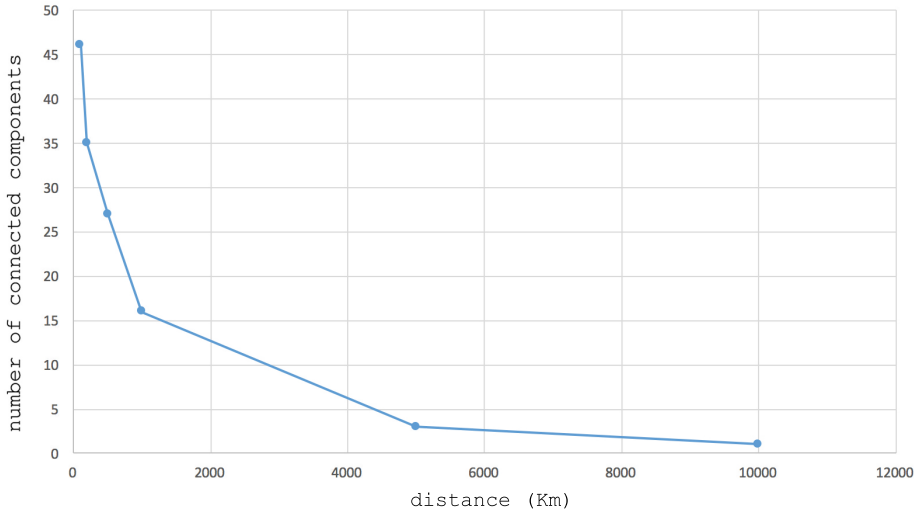


Fig. 9.2. Distribution of the number of connected components of the instances of our MIoT against distances

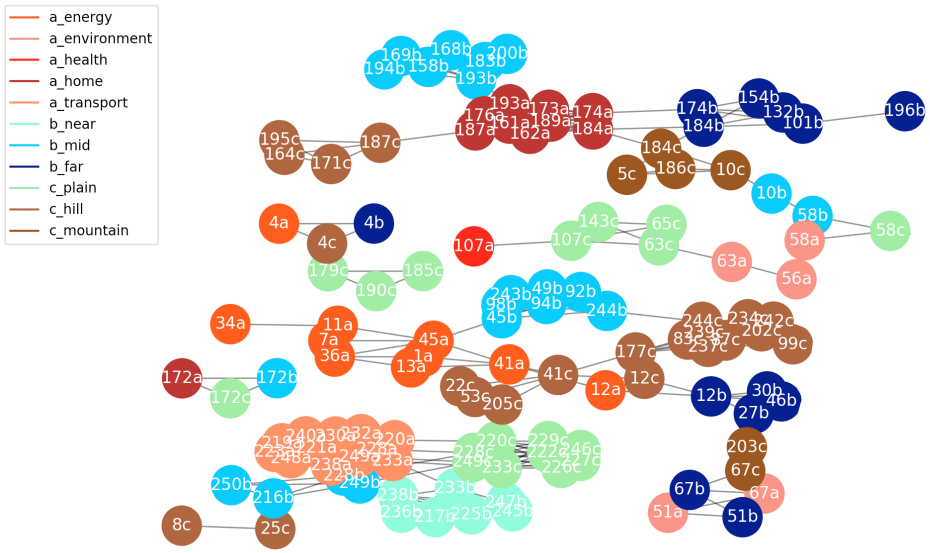


Fig. 9.3. Graphical representation of our MIoT

In this section, we present a case study aiming at comparing the classical vision of a unique global Internet of Things with the new MIoT-based vision of several Internets of Things connected to each other through cross nodes and cross edges. In our opinion, this case study can help the reader to be convinced of the practicality of the MIoT paradigm.

First, we must clarify that a slavish comparison between the previous vision of IoT and the MIoT-based vision is not possible, because this last paradigm associates more instances with the same object, one for each network joined by it. By contrast, the classical global IoT-based vision considers only objects and does not allow the

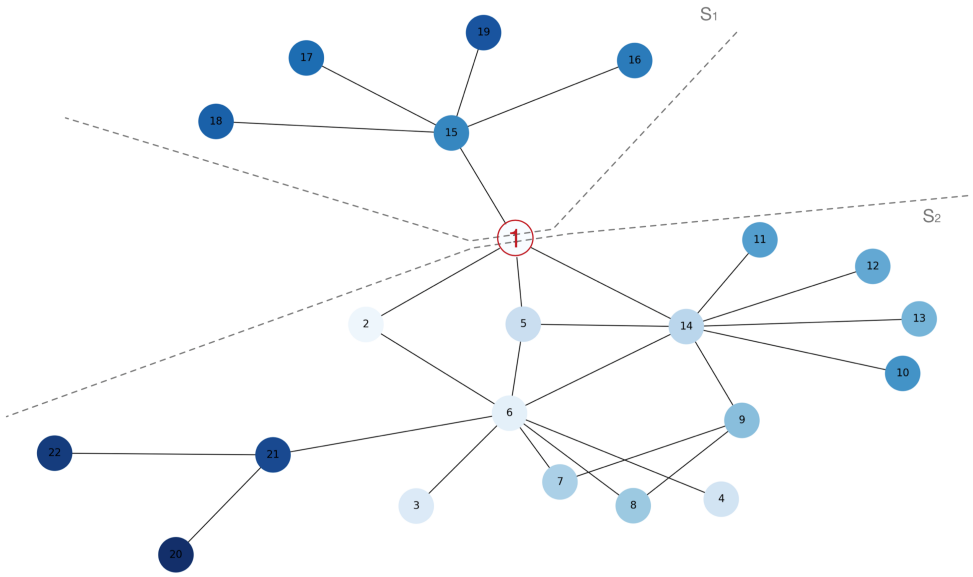


Fig. 9.4. Our case study

existence of more instances of the same object. In other words, the global IoT-based vision returns a coarser model of the involved things and their relationships, incapable of verifying if the same object shows different features or behaviors in different subnetworks of the global network. Vice versa, this verification is not only possible, but also natural, in the MIoT paradigm. Indeed, it is sufficient to investigate the different features and behaviors of the various instances of the same object in the IoTs they belong to.

After having made this important premise, which already represents a justification of the usefulness of the MIoT paradigm, we start by presenting our case study by which we aim at showing that the global IoT-based vision can provide imprecise information about the features and the roles of the corresponding things.

Since the global IoT-based vision does not consider object instances, in this case study we assume that all the instances of a cross object have been merged in a unique c-node.

With this considerations in mind, let us consider Figure 9.4. Here, we report a set of nodes each associated with an object. If we consider the global IoT-based vision, all these nodes form a unique IoT where it is possible to distinguish two quite separated subnetworks, called S_1 and S_2 in the figure, connected only thanks to the object represented by Node 1. If we consider the MIoT-based vision, we have two IoTs connected, by means of the object represented by Node 1, to form a MIoT.

Let us focus our attention on this node. Clearly, it is the most important node of this scenario because it is the only one allowing the communication and the co-

Nodes	*Betweenness Centrality*	*Degree Centrality*	*Closeness Centrality*	*Eigenvector Centrality*
1	0.39 (3)	0.19 (4)	0.44 (4)	0.30 (4)
2	0.07 (6)	0.09 (8)	0.41 (5)	0.20 (6)
3	0.00 (11)	0.05 (11)	0.33 ()	0.13 (14)
4	0.00 (12)	0.05 (12)	0.33 ()	0.13 (15)
5	0.07 (7)	0.14 (6)	0.47 (3)	0.34 (3)
6	0.52 (1)	0.38 (1)	0.48 (2)	0.34 (2)
7	0.01 (9)	0.09 (9)	0.34 ()	0.19 (7)
8	0.01 (10)	0.09 (10)	0.34 ()	0.19 (8)
9	0.04 (8)	0.14 (7)	0.37 (6)	0.23 (5)
10	0.0 (13)	0.04 (13)	0.35 (9)	0.13 (10)
11	0.0 (14)	0.04 (14)	0.35 (10)	0.13 (11)
12	0.0 (15)	0.04 (15)	0.35 (11)	0.13 (12)
13	0.0 (16)	0.04 (16)	0.35 (12)	0.13 (13)
14	0.48 (2)	0.38 (2)	0.52 (1)	0.49 (1)
15	0.35 (4)	0.23 (3)	0.35 (7)	0.11 (16)
16	0.0 (17)	0.05 (17)	0.26 (17)	0.03 (19)
17	0.0 (18)	0.05 (18)	0.26 (18)	0.03 (20)
18	0.0 (19)	0.05 (19)	0.26 (19)	0.03 (21)
19	0.0 (20)	0.05 (20)	0.26 (20)	0.03 (22)
20	0.0 (21)	0.05 (21)	0.26 (21)	0.04 (17)
21	0.18 (5)	0.14 (5)	0.35 (8)	0.15 (9)
22	0.0 (22)	0.05 (22)	0.26 (22)	0.04 (18)

Table 9.2. Betweenneess Centrality, Degree Centrality, Closeness Centrality and Eigenvector Centrality, and the corresponding ranks, for all the nodes of the case study of Figure 9.4

operation between the nodes of the subnetwork S_1 and the ones of the subnetwork S_2.

However, if we compute the classical centrality measures for the nodes of this network, we have that the rank of Node 1 is not very high in any centrality measure (see Table 9.2). In other words, if we adopt the global IoT-based vision, no centrality measure is capable of capturing the importance of this node. By contrast, the MIoT paradigm is capable alone of intrinsically evidencing the key role played by Node 1, without the need of computing any centrality measure.

With regard to this last observation, we are also aware that, in a real scenario, where the IoTs composing a MIoT are many and the number of c-objects is high, it could be extremely challenging to define a new MIoT-oriented centrality measure. This should be capable of determining the most relevant nodes in a MIoT taking also

(but not exclusively) into account if they are c-nodes or not. In the future, we plan to investigate the possibility to define such a measure.

9.4 CDS: a crawler tailored for MIoTs

9.4.1 Motivations underlying CDS

As pointed out in the Introduction, in real cases, when the number of involved things is huge, in order to investigate the main features of a MIoT and to extract useful knowledge from its data, a crawling strategy is mandatory. This strategy must be able to consider not only the instances and their connections in a single IoT (i.e., i-nodes and i-edges), but also the instances of the same objects (along with the corresponding connections) in different IoTs (i.e., c-nodes and c-edges). Furthermore, it must take into consideration that c-nodes and i-nodes have different nature and that c-nodes are more important than i-nodes in a MIoT, which implies that it must be possible to privilege c-nodes over i-nodes, if necessary. Finally, it must allow users to specify how much c-nodes must be privileged over i-nodes. Observe that this problem has a correspondence with the one of finding a crawler specifically tailored for a Social Internetworking Scenario and, therefore, a crawler privileging "me"-edges over intra-network edges and bridges over intra-network nodes.

In the past, several crawling strategies operating in a *single network* (and, therefore, in a *single IoT*) have been proposed. Among them, three very popular ones are Breadth First Search (BFS, for short), Random Walk (RW, for short) and Metropolis-Hastings Random Walk (MH, for short). BFS implements the classical Breadth First Search visit. RW selects the next node to visit uniformly at random among the neighbors of the current node. Both BFS and RW tend to favor power nodes (i.e., nodes having high outdegrees). As a consequence, both of them present bias in some network parameters [249]. MH is a more recent crawling strategy, conceived to unfavor power nodes in such a way as to remove the bias, in BFS and RW, caused by their tendency to favor this kind of node. It was shown that MH performs very well in a single network [171], especially for the estimation of the average degree of nodes. At each iteration, MH randomly selects a node n_j from the neighbors of the current node n_i. Then, it randomly generates a number p, belonging to the real interval $[0, 1]$. If $p \leq \frac{outdeg(n_i)}{outdeg(n_j)}$, where $outdeg(n_i)$ and $outdeg(n_j)$ are the outdegrees of n_i and n_j, it selects n_j as the new current node. Otherwise, it maintains n_i as the current node. The higher the outdegree of a node, the higher the probability that MH discards it. The way of proceeding of MH has been specifically conceived to reach the goal of disfavoring high-degree nodes in such a way as to remove the bias caused by them, as explained above.

In the past, BFS, RW and MH were deeply studied for single networks and it was found that none of them is always better than the other ones. However, no investigaton about the application of these strategies in a set of related IoTs (of which, SIoTs and MIoTs are specific cases) has been carried out. Thus, there is no evidence that they are still valid in this new context. Rather, it is easy to foresee that they will show some weaknesses, since they do not take into account the main actors of related IoTs, i.e., the instances of the same things in different IoTs and their connections (which represent c-nodes and c-edges in the MIoT paradigm).

We expect that these instances and their connections play a crucial role in crawling a set of related IoTs, since they allow different IoTs to be crossed, thus evidencing the main actors of related IoTs, i.e. c-nodes and c-edges, allowing their interconnections. These nodes and edges are not "standard" ones, due to their role. As shown in Section 9.3.2, we cannot see a set of related IoTs just as a unique huge IoT. By contrast, its nature, specificities and behavior must be strongly considered by a crawling strategy that aims to be effective and efficient for a set of related IoTs.

As it will be described in the next section, this original intuition has been fully confirmed by our experimental campaign, which clearly highlights the drawbacks of BFS, RW an MH when passing from a single IoT to a set of related IoTs.

9.4.2 Description of CDS

In the design of CDS, we start by analyzing some aspects limiting BFS, RW and MH in a set of related IoTs (and, therefore, also in a MIoT), in such a way as to overcome them.

BFS performs a Breadth First Search of a local neighborhood of the current node. Now, the average distance between two nodes of a single IoT is generally less than the one between two nodes of different IoTs. In fact, to pass from an IoT to another, it is necessary to cross a c-node and, since, in real cases, c-nodes are (much) less numerous than i-nodes, it could be necessary to generate a long path before reaching one of them. As a consequence, the local neighborhood considered by BFS includes one or a small number of IoTs.

To overcome this problem, a Depth First Search, instead of a BFS, could be performed. For this purpose, the way of proceeding of RW and MH should be included in our crawling strategy. However, since, generally, there is a limited number of c-nodes in an IoT, the simple choice to go in-depth blindly does not favor the crossing from an IoT to another. A solution that addresses the above issues could consist in the implementation of a "non-blind" Depth-First Search that favors c-nodes in the choice of the next node to visit. This is exactly the strategy we have chosen, and the name

we give to it, i.e., Cross Node Driven Search (CDS, for short), clearly reflects its way of proceeding.

Observe that this problem has a correspondence with the one of finding a crawler specifically tailored for a Social Internetworking Scenario and, therefore, a crawler privileging "me"-edges over intra-network edges and bridges over intra-network nodes.

However, following exactly the strategy mentioned previously would make it impossible to explore (at least partially) the neighborhood of the current node because the visit would proceed in-depth very quickly and, as soon as a c-node is encountered, there is a cross to another IoT. The overall result of this strategy would be an extremely fragmented crawled sample. To avoid this problem, given the current node, our crawling strategy explores a fraction of its neighbors before performing an in-depth search of the next node to visit.

To formalize our crawling strategy, we need to introduce the following parameters:

- *inf* (i-node neighbors fraction). It represents the fraction of the i-node neighbors of the current node that should be visited. It ranges in the real interval $(0, 1]$. When *inf* tends to 1, CDS behaves as BFS. By contrast, when *inf* tends to 0, CDS behaves as MH and RW[2]. In all these cases, CDS inherits all the strengths and the weaknesses of the corresponding strategies. Intermediate values of *inf*, suitably determined (see Section 9.4.3), allow CDS to maximize the pros and to minimize the cons of BFS, RW and MH.
- *cnf* (c-node neighbors fraction). It represents the fraction of the c-node neighbors of the current node that should be visited. It ranges in the real interval $(0, 1]$. It allows the tuning of the number of IoT crossings performed by CDS. The higher its value, the higher this number. Clearly, an excessive number of crossings could return a sample involving many IoTs of the MIoT but with a very little number of connections between each pair of IoTs. This could cause, in the Multiple-Network context, the same problem caused by RW in the Single-Network scenario. As a consequence, also for this parameter, a tradeoff is necessary.

For instance, in a configuration where $inf = 0.15$ and $cnf = 0.30$, CDS visits 15% of the i-node neighbors of the current node and 30% of the c-node neighbors of the current node.

We are now able to formalize our crawling strategy. We report its pseudocode in Algorithm 1.

[2] To be extremely accurate and precise, this is true if the parameter cnf (that we introduce below) is fixed to 1, in case we want to visit the whole MIoT, or to 0, in case we want to restrict our visit to just one IoT of the MIoT.

Algorithm 1 CDS

Notation We denote by $I(n)$ a function returning the number of i-node neighbors of the node n and by $C(n)$ a function returning the number of c-node neighbors of n.

Input $\mathcal{M}$: a MIoT composed of m IoTs; n_{it}: a non-negative integer; cnf, inf: a real number in the range [0,1]; $SeenNodes$, $VisitedNodes$, $VisitedCNodes$: a set of nodes

Output $SeenNodes$; $VisitedNodes$;

Variable v, w: a node

Variable p: a real number in the range [0,1]

Variable c: an integer number

Variable $NodeQueue$: a queue of nodes

1: $NodeQueue := \emptyset$
2: select a seed node s (not already present in $VisitedNodes$) from $\mathcal{M}$ uniformly at random
3: insert s in $NodeQueue$
4: **while** $n_{it} > 0$ **do**
5: extract a node v from $NodeQueue$
6: insert v in $VisitedNodes$
7: insert all the nodes adjacent to v in $SeenNodes$
8: **if** $(C(v) \geq 1)$ **then**
9: clear $NodeQueue$
10: $c := 0$
11: **while** $((c < \lceil cnf \cdot C(v) \rceil)$ **and** $(n_{it} > 0))$ **do**
12: let w be one c-node neighbor of v not in $VisitedCNodes$ selected uniformly at random
13: generate a number p in the real interval $[0, 1]$ uniformly at random
14: **if** $\left(p \leq \frac{C(v)+I(v)}{C(w)+I(w)}\right)$ **then**
15: insert w in $NodeQueue$ and in $VisitedCNodes$
16: $c := c + 1$
17: $n_{it} := n_{it} - 1$
18: **end if**
19: **end while**
20: **end if**
21: **if** $(I(v) \geq 1)$ **then**
22: $c := 0$
23: **while** $((c < \lceil inf \cdot I(v) \rceil)$ **and** $(n_{it} > 0))$ **do**
24: let w be one of the i-node neighbors of v selected uniformly at random
25: generate a number p in the real interval $[0, 1]$ uniformly at random
26: **if** $\left(p \leq \frac{I(v)}{I(w)}\right)$ **then**
27: insert w in $NodeQueue$
28: $c := c + 1$
29: $n_{it} := n_{it} - 1$
30: **end if**
31: **end while**
32: **end if**
33: **if** $((n_{it} > 0)$ **and** $(NodeQueue = \emptyset))$ **then**
34: **goto** 37
35: **end if**
36: **end while**
37: **if** $(n_{it} = 0)$ **then**
38: **return** $SeenNodes, VisitedNodes$
39: **else**
40: **return** $CDS(\mathcal{M}, n_{it}, cnf, inf, SeenNodes, VisitedNodes, VisitedCNodes)$
41: **end if**

CDS receives: *(i)* a MIoT $\mathcal{M}$, consisting of m IoTs; *(ii)* a non-negative integer n_{it}, denoting the number of iterations that must be still performed; *(iii)* cnf and inf; *(iv)* three sets of nodes, called $SeenNodes$, $VisitedNodes$ and $VisitedCNodes$, whose

semantics will be clear in the following. It returns $SeenNodes$ and $VisitedNodes$ after having updated them.

It exploits: *(i)* a function $I(n)$ returning the number of i-node neighbors of the node n; *(ii)* a function $C(n)$ returning the number of c-node neighbors of the node n; *(iii)* two support nodes v and w; *(iv)* a support real number p in the real interval $[0, 1]$; *(v)* a support counter c; *(vii)* a support queue $NodeQueue$ of nodes.

First CDS selects a seed node s (not already present in the list $VisitedNodes$ of the nodes already visited) from $\mathcal{M}$ uniformly at random, and inserts it in $NodeQueue$. Then, it starts a cycle that ends when the number n_{it} of iterations to be still performed is 0.

During each iteration, CDS extracts a node v from $NodeQueue$ and inserts it in $VisitedNodes$. At the same time, it inserts all the node neighbors of v in the list $SeenNodes$.

At this point, it computes $C(v)$ to verify if there exist c-node neighbors of v. In the affirmative case, it clears $NodeQueue$[3] and starts to examine these nodes until to either the number of examined c-nodes reaches the maximum value established through cnf or there are no available iterations.

During each of these internal iterations, CDS selects a node w, among the c-node neighbors of v not already present in the set $VisitedCNodes$ of the already visited c-nodes; the selection of w is performed uniformly at random. Then, it generates a real number p in the range $[0, 1]$ uniformly at random. If $p \leq \frac{C(v)+I(v)}{C(w)+I(w)}$, then w is inserted in both $NodeQueue$ and $VisitedCNodes$, c is increased of 1 and n_{it} is decreased of 1. Note that the last condition implements the strategy of MH into CDS, in such a way as to let CDS to inherit the pros of MH.

After having processed the c-node neighbors of v, CDS starts to process the i-node neighbors of v in an analogous way. In particular, it selects a node w among the i-node neighbors of v uniformly at random. Then, it generates a number p in the real interval $[0, 1]$ uniformly at random and, if $p \leq \frac{I(v)}{I(w)}$, it inserts w into $NodeQueue$, increases c of 1 and decreases n_{i_t} of 1.

CDS terminates the external cycle started at row 4 when $n_{it} = 0$ or when there are no nodes that can be visited starting from the current seed. In the former case, it returns $SeenNodes$ and $VisitedNodes$. In the latter case, it recursively calls another instance of itself in such a way as to re-start all the previous tasks from another seed node not already visited in the past.

[3] Observe that this task is performed to privilege c-nodes over i-nodes and to favor crossings from one IoT to another. Indeed, if $NodeQueue$ would have not been cleared, there was the risk to remain in the same IoT or, in any case, to visit a very small number of IoTs.

Iterations	10	20	30	40	50	60	70	80	90	100
Seen nodes	50	78	107	150	165	163	183	187	181	198
Visited nodes	11	21	34	48	59	68	94	102	105	125
IoT Crossings	4	9	14	17	24	24	33	40	30	43
Visited IoTs	5	6	7	9	9	9	10	10	10	10

Table 9.3. Number of seen nodes, number of visited nodes, number of IoT crossings and number of visited IoTs against the number of iterations performed by CDS

9.4.3 Experimental campaign

We carried out our experiments on the testbed presented in Section 9.3.1. In particular, we performed two kinds of experiment, namely:

- *setting of CDS*; in this case, we aimed to choose the most suitable values of the input parameters of CDS;
- *evaluation of CDS*; in this case, we compared CDS with BFS, RW and MH to quantitatively determine its strengths and weaknesses.

In the next subsections, we present each of these experiments.

Setting of CDS

As pointed out in Section 9.4.2, CDS needs three input parameters that can be used to make it more responsive to our needs. These parameters are: *(i)* inf, i.e. the i-node neighbors fraction that should be visited; *(ii)* cnf, i.e. the c-node neighbors fraction that should be visited; *(iii)* n_{it}, i.e. the maximum number of iterations.

We recall that our testbed consists of 315 nodes; 200 of them are i-nodes, whereas 115 of them are c-nodes.

First, we computed the variation of the number of seen and visited nodes, IoT crossings and visited IoTs against the variation of the number of performed iterations. Obtained results are reported in Table 9.3.

From the analysis of this table, we can see that:

- after 20 iterations, 24.76% of all nodes are seen, 6.67% of all nodes are visited and 54.55% of IoTs are visited;
- after 50 iterations, 52.38% of all nodes are seen, 18.73% of all nodes are visited and 81.81% of IoTs are visited;
- after 70 iterations, 58.10% of all nodes are seen, 29.84% of all nodes are visited and 90.91% of IoTs are visited;
- after 100 iterations, 62.85% of all nodes are seen, 39.68% of all nodes are visited and 90.91% of IoTs are visited.

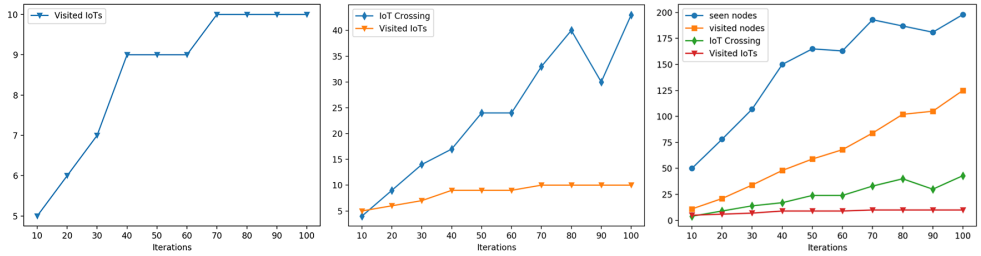


Fig. 9.5. Trends of the number of seen nodes, visited nodes, IoT crossings and visited IoTs against the number of iterations performed by CDS (trends are separated in the first two graphs and put together in the last one)

Taking into account these observations, as well as the trends of the corresponding measures reported in Figure 9.5, we observe that setting the number of iterations to 70 (or, more formally, setting $n_{it} = 0.22 \cdot |N|$) is a good tradeoff between the capability of sampling the highest possible number of the MIoT nodes and the effort required to perform this task.

After having set $n_{it} = 70$, we computed the variation of the number of seen and visited nodes, IoT crossings and, finally, visited IoTs against the variation of the values of inf and cnf. In particular, we considered five possible values of inf (i.e., $inf = 0$, $inf = 0.25$, $inf = 0.50$, $inf = 0.75$, and $inf = 1$) and five possible values of cnf (i.e., $cnf = 0$, $cnf = 0.25$, $cnf = 0.50$, $cnf = 0.75$, and $cnf = 1$). Obtained results are reported in Table 9.4.

From the analysis of this table, we can see that the best values for the four parameters are found when inf is low and cnf is high. This is totally in line with the semantics of these two coefficients, as well as with the role that they play in CDS. In particular, we observe that, if we consider the four parameters overall, the best pair of values is $inf = 0.25$ and $cnf = 0.75$.

Evaluation of CDS

In this experiment, we compared CDS with BFS, RW and MH. In this activity, the first preliminary task was to find reasonable metrics for evaluating the performances of crawlers that operate on a set of related IoTs. For this purpose, first we extended to the Multiple-Network context the metrics designed for evaluating the performances of crawlers that operate on a Single-Network context. Then, we introduced some other metrics specific for a set of related IoTs.

This section illustrates all our efforts in this direction and the results we have obtained. Specifically, it is organized in three subsections. The first presents our basic evaluation measures. The second describes a combined evaluation measure introduced

Seen nodes					
	$inf = 0$	$inf = 0.25$	$inf = 0.50$	$inf = 0.75$	$inf = 1$
$cnf = 0$	152	144	159	132	161
$cnf = 0.25$	200	178	201	212	171
$cnf = 0.50$	189	183	206	196	170
$cnf = 0.75$	199	212	204	172	204
$cnf = 1$	208	174	181	181	194
Visited nodes					
	$inf = 0$	$inf = 0.25$	$inf = 0.50$	$inf = 0.75$	$inf = 1$
$cnf = 0$	55	55	56	54	56
$cnf = 0.25$	64	61	65	65	62
$cnf = 0.50$	65	64	70	67	63
$cnf = 0.75$	71	70	69	62	70
$cnf = 1$	70	63	66	65	68
IoT crossing					
	$inf = 0$	$inf = 0.25$	$inf = 0.50$	$inf = 0.75$	$inf = 1$
$cnf = 0$	23	20	22	19	24
$cnf = 0.25$	29	26	31	31	26
$cnf = 0.50$	30	28	36	32	37
$cnf = 0.75$	36	37	34	26	35
$cnf = 1$	35	25	29	29	33
Visited IoTs					
	$inf = 0$	$inf = 0.25$	$inf = 0.50$	$inf = 0.75$	$inf = 1$
$cnf = 0$	9	8	8	8	10
$cnf = 0.25$	10	9	10	10	9
$cnf = 0.50$	9	9	10	9	9
$cnf = 0.75$	9	10	10	9	10
$cnf = 1$	10	9	9	9	9

Table 9.4. Number of seen nodes, visited nodes, IoT crossings and visited IoTs against the variation of inf and cnf

by us. Finally, the last presents the results of the test that we have performed by means of these measures.

Basic evaluation measures

The basic evaluation measures that we designed for our experimental campaign are the following:

- *Cross Node Ratio (CNR)*: This is a real number, in the interval $[0, 1]$, defined as the ratio of the number of crawled c-nodes to the number of all the c-nodes of the MIoT.
- *IoT Crossings (IC)*: This is a non-negative integer and denotes how many times the crawler switches from one IoT to another.

- *Visited IoTs (VI)*: This is a positive integer and measures how many different IoTs are visited by the crawler.
- *Unbalancing (UB)*: This is a non-negative real number defined as the standard deviation of the fraction of nodes discovered for each IoT w.r.t. the overall number of nodes discovered in the sample. UB ranges from 0, corresponding to the case in which each IoT is sampled with the same number of nodes, to a maximum value, corresponding to the case in which all sampled nodes belong to the same IoT.
- *Degree Bias (DB)*: This is a real number defined as the root mean squared error, for each IoT of the MIoT, of the average node degree estimated by the crawler and the one estimated by MH, which is considered the best crawling strategy for the estimation of the degree of a network node in the literature [249, 171]. If the crawled sample does not cover one or more IoTs, then these are not considered in the computation of DB.

If we consider the parallelism between MIoTs and Social Internetworking, we have that, in a Social Internetworking System: *(i)* CNR would return the ratio of the number of bridges discovered to the number of all the nodes in the sample; *(ii)* IC would measure how many times the crawler switches from one social network to another; *(iii)* VI would return how many different social networks are visited by the crawler; *(iv)* UB would represent the standard deviation of the percentages of nodes discovered for each social network w.r.t. the overall number of nodes discovered in the sample; *(v)* DB would denote the root mean squared error, for each social network of the SIS, of the average node degree estimated by the crawler and the one estimated by MH.

As for CNR, IC and VI, the higher their value, the higher the performance of the crawling strategy. By contrast, as far as UB and DB are concerned, the lower their values and the higher the performance of the crawling strategy. Observe that VI allows the evaluation of the crawler's capability of covering many IoTs of the MIoT. With regard to this measure, a further consideration is in order. Indeed, one could think that a fair crawling strategy should sample different IoTs proportionally to their respective overall size. Actually, this crawler behavior could result in incomplete samples in case of a high variance of these sizes. In fact, it could happen that some small IoTs would be not represented, or would be insufficiently represented, in the sample. CNR and IC are related to the coupling degree of the IoTs of the MIoT, whereas DB is related to the average degree.

A combined evaluation measure

Besides some separated metrics, each capturing an important aspect of the crawling strategy, it is certainly important to define a synthetic measure capable of capturing a sort of "overall" crawler behavior. Furthermore, this overall measure should allow users to tune the importance of the five metrics in it, which could be different in different application cases. A reasonable way to do this consists in defining the overall metric as a linear combination of the five ones introduced above, where the coefficients reflect the importance that users want to associate with them. We call *Overall Crawling Quality* (*OCQ*, for short) this measure and define it as:

$$OCQ = w_{CNR} \cdot \frac{CNR}{CNR_{max}} + w_{IC} \cdot \frac{IC}{IC_{max}} + w_{VI} \cdot \frac{VI}{VI_{max}} + w_{UB} \cdot (1 - \frac{UB}{UB_{max}}) + w_{DB} \cdot (1 - \frac{DB}{DB_{max}})$$

Here, CNR_{max}, IC_{max}, VI_{max}, UB_{max} and DB_{max} are the upper bounds of CNR, IC, VI, UB and DB, which, in a comparative experiment, can be set to the maximum value obtained by the crawlers into consideration. Furthermore, w_{CNR}, w_{IC}, w_{VI}, w_{UB} and w_{DB} are real numbers belonging to the interval $[0, 1]$ such that their overall sum is 1.

Before reasoning about the possible values of the five weights of OCQ, we point that the defined metrics are not completely independent of each other. In fact, if $CNR = 0$, then IC and VI are also 0. Furthermore, the value of CNR influences the values of both VI and UB. As a consequence, it is reasonable to assign different weights to the five metrics by associating the highest weights with the most influential ones. To perform this task, we defined an algorithm that is based on the Kahn's approach for topological sorting of graphs [224]. This algorithm uses a data structure called Metric Dependency Graph. This graph has a node n_i for each metric M_i; there exists an edge from n_i to n_j if the metric M_i influences the metric M_j. Each node has associated a weight. Initially all the node weights are set to 0.20 (see Figure 9.6). Our algorithm starts from a node with no outgoing edges and splits the corresponding weight (in equal parts) between itself and the nodes it depends on. Clearly, if a node has no incoming edge, it maintains its weight. After the split of the weight, our algorithm removes all the incoming edges from the corresponding nodes and repeats the previous tasks until all the nodes of the graph have been processed.

It is worth pointing out that the node processing order could be not unique, if there exists more than one node with no outgoing edges. However, it is possible to prove that the final metric weights returned by our algorithm do not depend on the adopted node processing order.

It is possible to formalize the previous algorithm in a closed formula allowing us to compute the weight w_i associated with each node n_i of the Metric Dependency Graph. In particular, we have:

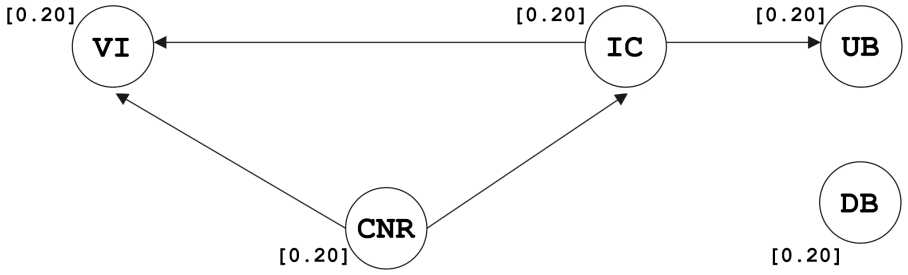


Fig. 9.6. Our Metric Dependency Graph

$$w_i = \frac{\frac{1}{1+indeg(n_i)} \cdot \left(w + \sum_{n_j \in OSet(n_j)} w_j\right)}{\sum_{k=1}^{5} w_k}$$

Here, $indeg(n_i)$ is the indegree of n_i, w is a number representing the initial weight of n_i (that, in our case, is 0.20 for all the five nodes) and $OSet(n_j)$ is the set of the nodes reachable from n_j through its outgoing edges. This formula indicates that w_i consists of two components; the former is the initial weight w; the latter represents the weight gained by n_i thanks to the fact that other nodes depend on it. In turn, n_i splits its weight among the nodes it depends on and itself; this is handled by the term $1 + indeg(n_i)$. The denominator of the formula is used to normalize w_i in the interval $[0, 1]$.

By applying the previous formula to our five metrics we obtained the following weight values: $w_{CNR} = 0.45$, $w_{IC} = 0.18$, $w_{VI} = 0.07$, $w_{UB} = 0.10$ and $w_{DB} = 0.20$.

Test results

We are now ready to analyze the performances of CDS, BFS, RW and MH when applied on a MIoT. For this activity, we used the testbed described in Section 9.3.1. We applied BFS, RW and MH to each MIoT by regarding it as a unique graph. Furthermore, in order to make the MIoT graph totally compliant with the inputs classically received by BFS, RW and MH, we considered a "condensed version" of the MIoT graph by putting just one node for each c-object. We run CDS with $inf = 0.25$ and $cnf = 0.75$, which, as pointed out in Section 9.4.3, are the parameter settings that guarantee the maximum number of IoT crossings. We report the obtained results in Table 9.5.

We recall that the higher the values of CNR, IC and VI and the lower the values of DB and UB, the better the performances of the strategies into examination.

From the analysis of Table 9.5, we can observe that, as far as CNR, IC, VI and UB are concerned, CDS outperforms BFS, RW and MH. For instance, the value of

	CDS	BFS	RW	MH
CNR	0.211	0.064	0.057	0.053
IC	9.133	6.400	2.333	2.333
VI	27.000	10.933	7.467	7,467
DB	3.476	0.844	0.026	0
UB	0.142	0.269	0.236	0.199

Table 9.5. Values of the five metrics obtained by CDS, BFS, RW and MH

CNR obtained by CDS is about 230% (resp., 273%, 296%) better than the one of BFS (resp., RW, MH).

The only metric for which CDS shows a worse performance than the other strategies is DB. In fact, as for this metric, the value obtained by MH is 0. This was expected because DB is measured having the value of MH as the reference one since, in the literature, it is well known that MH guarantees the best Degree Bias among all crawling strategies [249, 171]. BFS and RW obtain values of DB near to the ones of MH, whereas CDS shows the worst performance, even if it is still acceptable. The results obtained by CDS for DB were also expected because the purpose of this crawler is to privilege c-nodes over i-nodes. As a consequence, when a c-node is encountered, the node queue is cleared (see Line 4 in Algorithm 1) in such a way as to stimulate the IoT crossings and, ultimately, the visit of c-nodes, which is the main objective of our crawler. Clearing the node queue produces a distortion because several nodes directly connected to the current one will not be put in the set of visited nodes. In turn, this produces an effect in the degree bias and, ultimately, the worst performance of CDS, as far as the value of DB is concerned. However, observe that these results are obtained with the default configuration of CDS (i.e., $inf = 0.25$ and $cnf = 0.75$). Actually, if necessary, it is possible to configure CDS in such a way that it behaves as RW and MH, which present the best values of DB. In fact, as seen in Section 9.4.2, this behavior can be obtained by making inf tend to 0.

Since there is one parameter for which CDS shows the worst results w.r.t. the other three crawlers, it is particularly important the computation of the values of OCQ, because this parameter summarizes the overall performance of the crawlers into examination. We computed the values of OCQ for both the configuration that sets all the metric weights to 0.20 (we call it "Configuration A" in the following) and the one that takes the parameter dependencies into account ($w_{CNR} = 0.45$, $w_{IC} = 0.18$, $w_{VI} = 0.07$, $w_{UB} = 0.10$ and $w_{DB} = 0.20$ - we call it "Configuration B" in the following). In Table 9.6, we report the obtained results (we recall that the higher the value of OCQ and the better the performance of the corresponding crawler).

	CDS	BFS	RW	MH
Configuration A	0.695	0.433	0.383	0.409
Configuration B	0.747	0.410	0.399	0.407

Table 9.6. Values of *OCQ* obtained by CDS, BFS, RW and MH for the two weight configurations into examination

From the analysis of this table we can observe that, in both cases, CDS outperforms BFS, RW and MH. Interestingly, in the configuration taking the Metric Dependency Graph into account, CDS obtains even better results than in the other one.

In our opinion, these results clearly evidence that, in a MIoT scenario:

- The crawling strategies defined for single networks do not perform well because they do not consider the important differences existing between c-nodes and i-nodes and between c-edges and i-edges.
- A cross node centered crawler, like CDS, shows very satisfying results and, certainly, indicates a way to go for further crawler strategies specifically designed to operate on a set of related IoTs.

9.5 Analytical Discusssion

In this section, we propose an analytical discussion aiming at comparing our model and approach with other, more or less conventional, ones. We start by observing that, in the last years, the interest and the attention towards IoTs and sensor networks are enormously increased. This has led, and is currently leading, to a large variety of models and approaches. Some, very common and particularly interesting, families of approaches that can be recognized are the ones based on:

- fuzzy logic;
- neural networks;
- hierarchical models.

In the following, we present a comparison between our approach and each of these families.

Fuzzy logic based approaches allow the possibility that a thing belongs to more sets simultaneously [248, 357, 396, 28]. Also in our model, an object can belong to more IoTs, thanks to its instances. However, differently from fuzzy logic based approaches, in our case, when there is the instance of an object in an IoT, this means that the object surely belongs to that IoT. Instead, in fuzzy logic based approaches, an object belongs to a given IoT with a certain plausibility.

Neural network based approaches can exploit the potentialities of a highly dynamic structure, such as neural network [101, 378]. The dynamism of the support data structure certainly represents an analogy with our approach, which is based on an equally dynamic structure, i.e. social network. However, even if these two support data structures are graph based, they have totally different objectives. Indeed, neural networks are well suited for performing classifications and for handling non-linear scenarios. Social Networks are centered on node cooperation, node centralities and information diffusion. Furthermore, in a MIoT, there is no need to handle non-linearity.

Hierarchical approaches are certainly a bit more different from the MIoT paradigm than the other two families considered above [275, 440]. In fact, they mainly aim at detecting (more or less) hidden relationships among objects at different abstraction levels. Even if such a family of approaches is quite far from the current MIoT paradigm, it could represent a good starting point for an evolution of our model. Indeed, the current MIoT architecture consists of only two levels of control. Increasing the hierarchy length and, therefore, the granularity level, would allow the definition of more instances of one object in the same IoT, which could provide our model with a higher refinement capability.

Finally, to the best of our knowledge, the approach most similar to ours is the one described in [97]. In fact, analogously to what happens in a MIoT, in this approach an object is described by means of an ennuple. This choice allows an ordered representation of an object, its activities and its instances. However, very differently from our approach, the one of [97] models data coming from an IoT as a big data stream. This forces a kind of sampling allowing only the registration of the probability that a given object is in a given condition or in a given place. Interestingly, the approach of [97] provides the user with a strong support for data cleaning and integration. Instead, the MIoT paradigm does not address this issue because it assumes that cleaning and integration tasks have been performed before the construction of the MIoT graph.

10

Building Virtual IoTs in a Multiple IoTs scenario

10.1 Introduction

The Internet of Things (hereafter, IoT) is currently considered the new frontier of the Internet. As a matter of fact, a lot of research results, along with the continuous emergence of increasingly challenging issues to address, can be found in the literature [183, 397, 131, 354, 38, 180, 255].

One of the most effective ways to represent and handle the IoT scenario leverages social networking paradigm [35]. In this direction, several social network-based approaches to modeling and managing IoTs have been presented in the literature. Three of the most advanced ones are the SIoT (Social Internet of Things) [39, 150, 40, 394], the MIE (Multiple IoT Environment) [49] and the MIoT (Multiple IoTs) [50] paradigms. The MIoT paradigm is the last of these proposals; it aims at extending both SIoT and MIE in such a way as to preserve their strengths and avoid their weaknesses [50]. Roughly speaking, a MIoT can be seen as a set of related IoTs, i.e., as a set of related networks of things. Actually, a more precise definition of MIoT requires the introduction of the concept of instance of a thing in an IoT. Specifically, the instance of a thing in an IoT represents a virtual view of that thing in the IoT. The nodes associated with a thing in a MIoT represent the instances of the same thing in the different IoTs of the MIoT. Indeed, a thing can have several instances, one for each IoT which it participates to. The existence of more instances for one thing plays a key role in the MIoT paradigm because it allows the definition of cross relationships among the different IoTs.

We adopted the MIoT paradigm as the reference one in this chapter. There are several reasons which justify this choice. Indeed:

- The MIoT paradigm, like the SIoT and the MIE ones, introduces the idea that objects can show a social behavior in the environment where they operate. This feature allows several advantages, like the possibility of resource sharing (see [150, 40, 394] for a comprehensive idea of these advantages).

- Differently from SIoT, which introduces a social behavior of objects but still models IoT as one huge network of objects extended worldwide, MIE, and much more MIoT, allow the "breakdown" of the whole huge IoT into multiple networks of smart objects interconnected with each other. This way to proceed is analogous to the evolution of social networking into social internetworking [82]. In particular, MIoT allows the management of situations in which the same object shows different behaviors in different networks it joined. Furthermore, MIoT makes an object to act as a bridge between two objects allowing them to communicate even if they belong to different networks and, therefore, are not directly connected with each other.

Another important trend characterizing the current IoT scenario regards the existence of increasingly sophisticated and intelligent things. These are becoming increasingly smart and social, as well as more and more capable of performing computations and storage on their own. Furthermore, they are increasingly connected to each other through more and more complex and sophisticated frameworks, often based on cloud and edge computing [150, 40, 394]. The new smart and social capabilities of things and of the environments handling their interoperability paves the way to a sort of "humanization" of things, i.e., to apply to things concepts and ideas typically considered prerogative of humans. One of them is certainly the presence of a profile of a thing. Indeed, if a thing interacts with other things and exchange data with them, it is possible to determine what are the most common concepts handled by it and, based on them, to construct a corresponding profile. Analogously to the profile of a human, the one of a thing depends on its past behavior and on the profile of the other things with which it interacts. As a consequence, it could be possible to think about both a content-based and a collaborative-filtering approach to handling thing profiles.

Furthermore, starting from the real IoTs of a MIoT, it is possible to construct virtual communities of things, based on common interests. Once again, this is an attempt to transfer behaviors typical of humans to things. As a matter of fact, in Social Network Analysis, it is well recognized that, accordingly to the homophily concept [305, 408], humans tend to group together in communities sharing the same interests.

In the literature, a lot of efforts have been made to investigate human profiles and virtual communities of people, especially (but not only) in Social Network Analysis [400, 367]. Instead, these topics have been little investigated in the Internet of Things.

In this chapter, we aim at providing a contribution in this direction. First of all, we introduce the concept of profile of a thing. As the profile of a human, the one of a thing has two components. The former denotes its past behavior and can be used, for instance, to support content-based recommendations. The latter reflects its

neighbors, i.e., the other things with which it most frequently comes into contact; it can be exploited, for instance, to support collaborative filtering recommendations.

After this, we introduce the concept of topic-guided virtual IoTs in a MIoT and we propose two approaches (one supervised and one unsupervised) to the construction of them in a MIoT. Differently from the real IoTs of a MIoT, which may encompass things with very heterogeneous profiles, topic-guided virtual IoTs should include all and only those things whose profile refers to specific topics. The supervised approach requires a user to provide a set of keywords of her interest. It aims at constructing a thematic IoT comprising all the keywords specified by the user. If such an IoT does not exists, it returns more thematic IoTs that, in the whole, comprise all the keywords specified by the user. She can choose whether to accept this set of virtual IoTs or to modify her query. The unsupervised approach tries to partition a MIoT into a set of virtual IoTs characterized by the maximum internal cohesion (in terms of topics present in the profiles of the corresponding things) and the minimum external coupling. Virtual IoTs in a MIoT provide a logic representation of the objects of a MIoT, which is not based on real links but on the content exchanged by them. As will be clear in the following, this can favor the effectiveness of information exchange, the construction of communities of objects (and, possibly, of the corresponding users) sharing the same interests and the suggestions of the objects most adequate to a given exigency.

This chapter is organized as follows: in Section 10.2, we examine related literature. In Section 10.3, we provide an overview of the MIoT paradigm, because its comprehension is necessary to understand the rest of this chapter. In Section 10.4, we introduce our definition of a thing's profile. In Section 10.5, we propose our approaches to construct topic-guided virtual IoTs in a MIoT. In Section 10.6, we present our tests devoted to verify the performance of our approach.

10.2 Related Literature

Since its introduction some years ago, the term "Internet of Things - IoT" has been associated with a huge variety of concepts, technologies and solutions [38, 41, 314, 363]. In the latest years, with the advent of new technologies, such as big data and social networking, the very definition of this term is continuously changing. What IoT will become in the future depends on the evolution of these technologies [438] and their interaction with several other ones, such as Information Centric Networks [425, 479, 480, 32, 372, 33, 361] and Cloud [131, 433, 229]. As a matter of fact, the strengths of these last ones are exactly the features necessary to overcome the weaknesses of the

current IoT concept [467]. Some examples of this combination can be already found in the literature [150, 180, 449, 448].

The first attempts to apply social networking to the IoT domain can be found in [182, 333, 245, 205]. In these papers, the authors propose to use human social network relationships to share services provided by a set of things. An important step forward is performed in [39], where the SIoT paradigm is introduced. Here, the authors propose an approach to creating relationships among things, without requiring the owner intervention. Thanks to this idea, things can autonomously crawl the network to find services and resources of their interest provided by other things. In [42], the same authors clearly highlight what are the main strengths of SIoT. Specifically: *(i)* the SIoT structure can be dynamically modified to ensure network navigability and to find new things; *(ii)* scalability is guaranteed, like in human social networks; *(iii)* a level of trustworthiness among things can be established; *(iv)* the past social network approaches can be redefined to solve problems typical of the IoT context [342].

One of the major drawbacks of the current IoT scenario is the presence of different technologies and solutions proposed by independent vendors to enable networking among objects. This poses the basis to a subsequent set of issues ranging from concept matching to technical compatibility, if heterogeneous smart-object-network solutions should be involved in the creation of a unique interoperable IoT [336, 413]. In this research context, different works partially addressing and solving these problems have been proposed. Specifically, [166] presents a study on how ontologies and semantic data processing can be used to improve interoperability across heterogeneous IoT platforms. The authors consider two use cases, namely *Health Care* and *Trasportation and Logistics*, and, for each of them, provide a survey on the main ontologies available to describe and generalize concepts and relations.

In [265], instead, the authors focus their attention on the definition of a new framework for a fully functional mobile ad-hoc social network. In this chapter, the term "mobile ad-hoc social network" refers to an IoT made of mobile devices. Of course, communication between this type of objects may happen in such a wide range of modes so that the referring scenario can be considered as a constellation of mobile networks interacting with each other. Concepts from real social networks are borrowed to define user profiles, which are built starting from the objects they own and the social network they belong to. One of the main contributions of this proposal is the definition of a profile-matching strategy based on semantics.

Another contribution in the context of interoperability is the one proposed in [428]. Here, the authors illustrate a novel architecture in which objects interact with each other by leveraging an open source cloud platform. The interaction among smart devices is information-and-service-driven and can be performed in both a centralized

and a peer-to-peer mode. In [481], the authors propose *Acrost*, a system capable of retrieving data spread among heterogeneous IoT platforms by leveraging topics and semantics awareness. To build the metadata, Acrost uses two methodologies: the former exploits regular expression-based approaches, whereas the latter makes use of random fields-based strategies.

In order to address the issues arising when the interoperability among heterogeneous IoTs must be guaranteed, another research line proposes the extension of the results concerning Social Internetworking [82, 335] (instead of social networking) to the Internet of Things. By following this strategy, the MIE (Multiple IoT Environment) [49] and the MIoT (Multiple IoTs) [50] paradigms have been proposed. As specified in the Introduction, this last paradigm is the reference one for this chapter.

In [130], the authors present an approach to constructing a virtual data mart on which several knowledge discovery tasks can be performed. Clearly the kinds of virtual source constructed in the approach of [130] and in our own are very different. However, the general ideas underlying the two approaches are similar.

In the past, a lot of efforts have been made to investigate human profiles and virtual communities of people, especially (but not only) in Social Network Analysis ([400, 367] provide two surveys about these topics). Instead, these issues have been little investigated in the Internet of Things. Specifically, to the best of our knowledge, a comprehensive, high-level abstraction approach to building and managing a profile of a thing, which also takes into account the content it exchanges during its interactions with other things, has not yet been proposed. Instead, some approaches focusing on community detection in IoT have been presented in the very recent literature. Even if they are very different (both in their purposes and in their ways proceed) from the ones of our approach, in the following we present an overview of some of them.

The approach of [453] uses structural information derived from the complex graph of an IoT to extract communities. It exploits a neighbor-based strategy to detect also overlapping communities. The approach of [230] uses data produced by sensors to define a multi-dimensional clustering. The obtained clusters are then mapped to communities of nodes in the original IoT network. To cope with the size of the data graph, the authors leverage state-of-the-art community detection approaches. Finally, they present a new community detection approach that enhances the Girvan-Newman algorithm by using hyperbolic network embedding.

Other works, instead, use knowledge from social networks to refine their results. As an example, [317] proposes a community definition strategy combining both IoT information and structural data coming from the social network (relationship among users), which object owners belong to. This approach does not consider semantics and contents, but leverages only network structure. A similar method is proposed in [55],

even though here the strategy works in the opposite way. In fact, first communities are derived from structural information of owners' social networks and, then, objects are seen as resources available inside each community.

Finally, the authors of [244] propose a new community detection algorithm working in a Social Internet of Things (SIoT) scenario. To achieve their objective, they make use of three metrics, namely social similarity, preference similarity and movement similarity. Social similarity is defined according to the concept of cooperativeness and community interest proposed in [334]. Preference similarity takes into account resource and service preferences of the involved things in the network. Finally, movement similarity specifies how much and how long two or more nodes are spatially close.

In [316], the authors propose a community detection approach working on an architecture capable of integrating the Internet of Things and social networking. This approach assumes that two nodes belong to the same community only if they are at most one hop apart and have at least two mutual friends. In order to construct communities, it exploits graph mining techniques.

10.3 The MIoT paradigm

In this section, we provide an overview of the MIoT paradigm, described in detail in [50], because it is the reference one for our definitions of virtual IoTs in a MIoT.

A MIoT $\mathcal{M}$ consists of a set of m Internets of Things. Formally speaking:

$$\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\} \tag{10.1}$$

where $\mathcal{I}_k$ is an IoT.

Let o_j be an object of $\mathcal{M}$. We assume that, if o_j belongs to $\mathcal{I}_k$, it has an instance ι_{j_k}, representing it in $\mathcal{I}_k$. The instance ι_{j_k} consists of a virtual view (or, better, a virtual agent) representing o_j in $\mathcal{I}_k$. For example, it provides all the other instances of $\mathcal{I}_k$, and the users who interact with $\mathcal{I}_k$, with all the necessary information about o_j. Information stored in ι_{j_k} is represented according to the format and the conventions adopted in $\mathcal{I}_k$.

A MIoT $\mathcal{M}$ can be represented by means of a graph-based notation. In particular, each IoT $\mathcal{I}_k \in \mathcal{M}$ can be modeled by means of a graph $G_k = \langle N_k, A_k \rangle$. In this case:

- N_k is the set of the nodes of G_k; there is a node n_{j_k} for each instance $\iota_{j_k} \in \mathcal{I}_k$, and vice versa.
- A_k is the set of the arcs of G_k; there is an arc $a_{jq_k} = (n_{j_k}, n_{q_k})$ if there exists a physical link from n_{j_k} to n_{q_k}.

Finally:

$$\mathcal{M} = \langle N, A \rangle \tag{10.2}$$

Here:

$$N = \bigcup_{k=1}^{m} N_k; \tag{10.3}$$

$$A = A_I \cup A_C, \tag{10.4}$$

where

$$A_I = \bigcup_{k=1}^{m} A_k \tag{10.5}$$

and

$$A_C = \{(n_{j_k}, n_{j_q}) | n_{j_k} \in N_k, n_{j_q} \in N_q, k \neq q\}. \tag{10.6}$$

A_I is the set of the inner arcs (hereafter, *i-arcs*) of $\mathcal{M}$; they relate instances (of different objects) belonging to the same IoT. A_C is the set of the cross arcs (hereafter, *c-arcs*) of $\mathcal{M}$; they relate instances of the same object belonging to different IoTs.

The description of the MIoT paradigm presented above highlights that it is possible to model a MIoT at two abstraction levels. The former represents a MIoT as a network and exploits concepts typical of this environment (such as nodes, arcs and so on). The latter models a MIoT as a set of IoTs and makes use of concepts closer to this scenario (such as instances, objects and so forth). Clearly, these two representations are simply two viewpoints of the same environment, and the concepts adopted by them can be used interchangeably. For example, there is a biunivocal correspondence between a node and an instance. However, in the reality, there are some cases in which it is better to use the concept of a node (for example, when we discuss about paths in a network - see below), whereas there are other situations in which it is better the use of the concept of instance (for example, when we discuss about the transactions carried out by two smart objects).

Furthermore, in a MIoT context, *a set MD_j of metadata* can be associated with an object o_j. Our metadata model refers to the one of the IPSO (Internet Protocol for Smart Objects) Alliance [5]. Specifically MD_j consists of three subsets, namely: *(i)* MD_j^D, i.e., the set of *descriptive metadata*; *(ii)* MD_j^T, i.e., the set of *technical metadata*; *(iii)* MD_j^B, i.e., the set of *behavioral metadata*. All details about these metadata can be found in [50].

10.4 Definition of a thing's profile

In this section, we present our definition of a thing's profile, which represents a first important contribution of this chapter. As pointed out in the Introduction, analogously to what happens for human profiles, the profile of a thing can have two components. The former registers its past behavior and is extremely useful for content-based recommendations; for this reason, we call it "content-based component" in the following. The latter registers the main features of those things with which it mostly interacted in the past and can be used for collaborative filtering recommendations; for this reason, we call it "collaborative filtering component" in the following.

Before illustrating in detail the profile of a thing, we must introduce some preliminary concepts. First of all, given two instances ι_{j_k} of o_j and ι_{q_k} of o_q in $\mathcal{I}_k$, we can define the set $tranSet_{jq_k}$ of the transactions from ι_{j_k} to ι_{q_k} as follows:

$$tranSet_{jq_k} = \{T_{jq_{k_1}}, T_{jq_{k_2}}, \cdots, T_{jq_{k_v}}\} \tag{10.7}$$

A transaction $T_{jq_{k_t}} \in tranSet_{jq_k}$ is represented as:

$$T_{jq_{k_t}} = \langle reason_{jq_{k_t}}, source_{jq_{k_t}}, dest_{jq_{k_t}}, start_{jq_{k_t}}, finish_{jq_{k_t}}, success_{jq_{k_t}}, content_{jq_{k_t}} \rangle \tag{10.8}$$

Here:

- $reason_{jq_{k_t}}$ denotes the reason why $T_{jq_{k_t}}$ occurred, chosen among a set of predefined values.
- $source_{jq_{k_t}}$ indicates the starting node of the path followed by $T_{jq_{k_t}}$.
- $dest_{jq_{k_t}}$ represents the final node of the path followed by $T_{jq_{k_t}}$.
- $start_{jq_{k_t}}$ denotes the starting timestamp of $T_{jq_{k_t}}$.
- $finish_{jq_{k_t}}$ indicates the ending timestamp of $T_{jq_{k_t}}$.
- $success_{jq_{k_t}}$ denotes whether $T_{jq_{k_t}}$ was successful or not; it is set to `true` in the affirmative case, to `false` in the negative one, and to `NULL` if $T_{jq_{k_t}}$ is still in progress.
- $content_{jq_{k_t}}$ indicates the content "exchanged" from ι_{j_k} to ι_{q_k} during $T_{jq_{k_t}}$. In its turn, $content_{jq_{k_t}}$ presents the following structure:

$$content_{jq_{k_t}} = \langle format_{jq_{k_t}}, fileName_{jq_{k_t}}, size_{jq_{k_t}}, topics_{jq_{k_t}} \rangle \tag{10.9}$$

Here:

  - $format_{jq_{k_t}}$ indicates the format of the content exchanged during $T_{jq_{k_t}}$; the possible values are: "audio", "video", "image" and "text".
  - $fileName_{jq_{k_t}}$ denotes the name of the transmitted file.

- $size_{jq_{k_t}}$ indicates the size in bytes of the content.
- $topics_{jq_{k_t}}$ indicates the set of the content topics; it consists of a set of keywords representing the subjects exchanged during $T_{jq_{k_t}}$. It can be formalized as: $topics_{jq_{k_t}} = \{(kw^1_{jq_{k_t}}, nkw^1_{jq_{k_t}}), (kw^2_{jq_{k_t}}, nkw^2_{jq_{k_t}}), \ldots, (kw^w_{jq_{k_t}}, nkw^w_{jq_{k_t}})\}$. In other words, the set of the topics of the t^{th} transaction from ι_{j_k} to ι_{q_k} consists of w pairs; each pair consists of a keyword and the corresponding number of occurrences.

Now, we can define the set $tranSet_{j_k}$ of the transactions performed by ι_{j_k} in $\mathcal{I}_k$. Specifically, let $Inst_k$ be the set of the instances of $\mathcal{I}_k$. Then:

$$tranSet_{j_k} = \bigcup_{\iota_{q_k} \in Inst_k, \iota_{q_k} \neq \iota_{j_k}} tranSet_{jq_k} \tag{10.10}$$

In other words, the set $tranSet_{j_k}$ of the transactions performed by an instance ι_{j_k} is given by the union of the sets of the transactions from ι_{j_k} to all the other instances of $\mathcal{I}_k$.

After having defined $tranSet_{j_k}$, we must introduce the following operators:

- $\biguplus$: it receives a set $\{entitySet_1, entitySet_2, \cdots, entitySet_t\}$ of entity sets and performs their union not eliminating the duplicates but reporting the number of their occurrences. Therefore, this operator returns a set of pairs $\{(entity_1, ne_1), (entity_2, ne_2), \cdots, (entity_w, ne_w)\}$ in which the pair $(entity_r, ne_r)$ indicates the r^{th} entity and the number of its occurrences. In counting it, $\biguplus$ takes the presence of synonymies and homonymies into account. These properties can be computed (for terms, images, etc.) by applying the classical approaches proposed in the past literature [62, 124].
- $avgFileSize$: it receives a set of files and computes their average size.

We are now able to define the profile $\mathcal{P}_{jq_k}$ of the relationship existing between two instances ι_{j_k} and ι_{q_k}, which performed a set $tranSet_{jq_k} = \{T_{jq_{k_1}}, T_{jq_{k_2}}, \cdots, T_{jq_{k_v}}\}$ of transactions. As we will see in the following, this profile plays a crucial role in the definition of the content-based component of a thing's profile and is indirectly used also in the definition of the collaborative filtering component of it. Specifically:

$$\begin{aligned}\mathcal{P}_{jq_k} = \langle & reasonSet_{jq_k}, sourceSet_{jq_k}, destSet_{jq_k}, avgSzAudio_{jq_k}, avgSzVideo_{jq_k}, \\ & avgSzImage_{jq_k}, avgSzText_{jq_k}, successFraction_{jq_k}, topicSet_{jq_k}\rangle\end{aligned} \tag{10.11}$$

where:

- $reasonSet_{jq_k} = \biguplus_{t=1..v}(reason_{jq_{k_t}})$;
- $sourceSet_{jq_k} = \biguplus_{t=1..v}(source_{jq_{k_t}})$;

- $destSet_{jq_k} = \biguplus_{t=1..v}(dest_{jq_{k_t}})$;
- $avgSzAudio_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}audio\text{”}\}$;
- $avgSzVideo_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}video\text{”}\}$;
- $avgSzImage_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}image\text{”}\}$;
- $avgSzText_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}text\text{”}\}$;
- $successFraction_{jq_k} = \frac{|\{T_{jq_{k_t}} | T_{jq_{k_t}} \in tranSet_{jq_k}, success_{jq_{k_t}} = true\}|}{v}$;
- $topicSet_{jq_k} = \biguplus_{t=1..v}(topics_{jq_{k_t}})$.

If we introduce the operator $\bigsqcup$, which compactly represents the set of operations for obtaining a profile of a pair of instances $\mathcal{P}_{jq_k}$ starting from the corresponding transactions, we can formalize the previous tasks by means of only one operation as follows:

$$\mathcal{P}_{jq_k} = \bigsqcup_{t=1..v} T_{jq_{k_t}} \tag{10.12}$$

Now, let ι_{j_k} be the instance of the object o_j in the IoT $\mathcal{I}_k$. Let $Inst_{j_k}$ be the set of the instances of $\mathcal{I}_k$ with which ι_{j_k} performed at least one transaction in the past. In this case, we can define the content-based component of the profile $\mathcal{P}_{j_k}$ of ι_{j_k} as:

$$\mathcal{P}_{j_k} = \bigsqcup_{\iota_{q_k} \in Inst_{j_k}} \mathcal{P}_{jq_k} \tag{10.13}$$

Finally, let o_j be an object and let $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_l\}$ be the set of the IoTs which it participates to. Let $ObjInst_j$ be the instances of o_j in the IoTs of the MIoT. We can define the content-based component of the profile $\mathcal{P}_j$ of o_j as:

$$\mathcal{P}_j = \bigsqcup_{\iota_{j_k} \in ObjInst_j} \mathcal{P}_{j_k} \tag{10.14}$$

After having defined the content-based component of an instance and an object, in order to present the corresponding collaborative filtering components, we must introduce the concept of neighborhoods of an instance ι_{j_k} in an IoT $\mathcal{I}_k$. Specifically, the structural neighborhood $sNbh(\iota_{j_k})$ of ι_{j_k} is defined as:

$$sNbh(\iota_{j_k}) = sNbh^{out}(\iota_{j_k}) \cup sNbh^{in}(\iota_{j_k}) \tag{10.15}$$

where:

$$sNbh^{out}(\iota_{j_k}) = \{\iota_{q_k} | (n_{j_k}, n_{q_k}) \in A_I\} \tag{10.16}$$

$$sNbh^{in}(\iota_{j_k}) = \{\iota_{q_k} | (n_{q_k}, n_{j_k}) \in A_I\} \tag{10.17}$$

Furthermore, we can also define the behavioral neighborhood $bNbh(\iota_{j_k})$ of ι_{j_k} as:

$$bNbh(\iota_{j_k}) = bNbh^{out}(\iota_{j_k}) \cup bNbh^{in}(\iota_{j_k}) \tag{10.18}$$

where:

$$bNbh^{out}(\iota_{j_k}) = \{\iota_{q_k} | \iota_{q_k} \in sNbh^{out}(\iota_{j_k}), |tranSet_{jq_k}| > 0\} \tag{10.19}$$

$$bNbh^{in}(\iota_{j_k}) = \{\iota_{q_k} | \iota_{q_k} \in sNbh^{in}(\iota_{j_k}), |tranSet_{qj_k}| > 0\} \tag{10.20}$$

In other words, $bNbh(\iota_{j_k})$ consists of those instances directly connected to ι_{j_k} from the structural viewpoint that shared at least one transaction with ι_{j_k}.

We are now able to present the collaborative filtering component $\mathcal{P}'_{j_k}$ of the profile of an instance ι_{j_k} in $\mathcal{I}_k$. It can be defined as follows:

$$\mathcal{P}'_{j_k} = \bigsqcup_{\iota_{q_k} \in bNbh(\iota_{j_k})} (\mathcal{P}_{q_k} \sqcup \mathcal{P'}_{q_k}) \tag{10.21}$$

Clearly, this definition is recursive and an accurate computation would require the resolution of a system with a number of equations and variables equal to the number of instances. In real situations, as there could be thousands or millions of instances in a MIoT, the time necessary to solve this system may easily become unacceptable. As a consequence, it appears reasonable to consider an approximate definition of $\mathcal{P}_{q_k}$ that is much simpler to handle. It is formalized as:

$$\mathcal{P}'_{j_k} = \bigsqcup_{\iota_{q_k} \in bNbh(\iota_{j_k})} \mathcal{P}_{q_k} \tag{10.22}$$

After having introduced the two components of the profile of an instance ι_{j_k} of $\mathcal{I}_k$, we can combine them for defining the overall profile $\overline{\mathcal{P}_{j_k}}$ of ι_{j_k}. It is defined as the union of the profiles $\mathcal{P}_{j_k}$ and $\mathcal{P}'_{j_k}$ performed by means of the operator $\sqcup$:

$$\overline{\mathcal{P}_{j_k}} = \mathcal{P}_{j_k} \sqcup \mathcal{P}'_{j_k} \tag{10.23}$$

Finally, we can define the overall profile of an object o_j as follows:

$$\overline{\mathcal{P}_j} = \bigsqcup_{k=1..l} \overline{\mathcal{P}_{j_k}} \tag{10.24}$$

10.5 Topic-guided virtual IoTs in a MIoT and approaches to constructing them

In this section, we present a supervised and an unsupervised approach to constructing topic-guided virtual IoTs in a MIoT.

10.5.1 Supervised approach

The supervised approach for the construction of topic-guided virtual IoTs in a MIoT requires the user to specify a query Q consisting of some keywords of her interest. It tries to construct a thematic virtual IoT in such a way that each of its instances contains at least one keyword of Q in the content-based component of its profile. If such a virtual IoT does not exist, our approach returns a minimal set of thematic IoTs that, on the whole, contain, in the content-based component of the profile of their instances, all the keywords specified by the user. In this last case, she can choose whether to accept this set of IoTs or modify her query.

Before describing in detail this approach, we must introduce a new operator J^* that represents a modified Jaccard coefficient, as we will see below.

J^* receives two sets of topics[1] $topicSet = \{(kw_1, nkw_1), (kw_2, nkw_2), \cdots, (kw_p, nkw_p)\}$ and $topicSet' = \{(kw'_1, nkw'_1), (kw'_2, nkw'_2), \cdots,$
$(kw'_p, nkw'_p)\}$ and computes the Jaccard coefficient between them. In carrying out this task, it considers the number of occurrences of each keyword and its possible synonyms.

More formally, first it computes the set:

$$commonTS = \{(kw, nkw + nkw') | (kw, nkw) \in topicSet, (kw', nkw') \in topicSet', \\ kw \text{ is identical to or synonymous of } kw'\} \tag{10.25}$$

Then, it computes the final result as:

$$J^*(topicSet, topicSet') = \frac{\sum_{(kw,nkw) \in commonTS} nkw}{\sum_{(kw,nkw) \in topicSet} nkw + \sum_{(kw',nkw') \in topicSet'} nkw'} \tag{10.26}$$

After having introduced J^*, we can describe our approach. Specifically:

- It starts when a user specifies a query Q consisting of r keywords:

$$Q = \{kw_1, kw_2, \cdots, kw_r\} \tag{10.27}$$

  It searches for all the instances of the MIoT having at least one topic whose keyword is identical to, or synonymous of, at least one keyword specified in Q. These instances, as a whole, represent the set of candidate instances to be included in the new thematic view. We call this set $\mathcal{CI}$ (Candidate Instances).

[1] We recall that, in our context, a topic is a pair (kw, nkw), where kw is a keyword and nkw is the corresponding number of occurrences.

- However, the fact that an instance $\iota \in \mathcal{CI}$ has a keyword in common with Q is necessary but not sufficient for it to be chosen. In fact, it is advisable that ι has more keywords in common with Q and, possibly, that the common keywords are among the ones of ι with the highest number of occurrences. This condition can be guaranteed by the usage of the operator J^*.
  In particular, our approach first constructs $Q' = \{(kw, 1) | kw \in Q\}$ in such a way as to make the application of J^* on the keywords specified by the user possible. Then, it constructs the set $\mathcal{RI}$ (Real Instances) of those instances of $\mathcal{CI}$ whose topics have a significant similarity with the keywords of Q:

$$\mathcal{RI} = \{\iota \in \mathcal{CI} | J^*(topicSet_\iota, Q') > th_J\} \tag{10.28}$$

  Here, th_J is a suitable tuning threshold.
- Now, our approach can start to construct the thematic view $\mathcal{V}_Q$ corresponding to Q.
  - It first creates a node n_ι in $\mathcal{V}_Q$ for each instance ι of $\mathcal{RI}$. Let n_{ι_1} and n_{ι_2} be the nodes corresponding to two instances ι_1 and ι_2 belonging to $\mathcal{RI}$.
    - If an i-arc exists between the nodes corresponding to ι_1 and ι_2 in the MIoT $\mathcal{M}$, then an i-arc is also created between the nodes n_{ι_1} and n_{ι_2} in $\mathcal{V}_Q$.
    - Instead, if a c-arc exists between the nodes corresponding to ι_1 and ι_2 in $\mathcal{M}$, then n_{ι_1} and n_{ι_2} are merged in a unique node $n_{\iota_{12}}$ in $\mathcal{V}_Q$. This task is motivated by the fact that n_{ι_1} and n_{ι_2} represent different instances of the same object in different real IoTs, but they represent the same instance in the same virtual IoT; as a consequence, they must be merged and no cross arc can exist between them. The profile $\overline{\mathcal{P}_{12}}$ of $n_{\iota_{12}}$ is obtained by applying the operator $\bigsqcup$ on the profiles $\overline{\mathcal{P}_1}$ of ι_1 and $\overline{\mathcal{P}_2}$ of ι_2.
- Finally, our approach adds a disconnected node in $\mathcal{V}_Q$ for each keyword in Q such that there is no MIoT instance having at least one topic whose keyword is identical to, or synonymous of, it[2].
- At this point, two cases may occur. In particular:
  - It could happen that $\mathcal{V}_Q$ is connected. In this case, it is returned as the answer to the query Q submitted by the user.
  - If $\mathcal{V}_Q$ is not connected and if the number of its connected components is less than a certain threshold, our approach adds the minimum number of "fictitious" i-arcs necessary to make $\mathcal{V}_Q$ connected.
  - Otherwise, if the number of connected components of $\mathcal{V}_Q$ is higher than a certain threshold, our approach concludes that a unique thematic virtual IoT

[2] The rationale underlying this step will be clearer in the following.

corresponding to the keywords specified by the user does not exist and returns the thematic views related to the connected components of $\mathcal{V}_Q$. At this point, the user can decide whether to accept these thematic views or to modify the query in such a way as to construct a unique thematic view by re-applying all the above mentioned steps starting from the new query.

10.5.2 Unsupervised approach

The unsupervised approach begins with the construction of a support network $\mathcal{N}$ starting from the MIoT $\mathcal{M}$. In particular:

- For each node n_{ι_k} of $\mathcal{M}$, a node $\overline{n_{\iota_k}}$ is added in $\mathcal{N}$.
- For each i-arc $(n_{\iota_{j_k}}, n_{\iota_{q_k}})$ in $\mathcal{M}$, an (unoriented) arc $(\overline{n_{\iota_{j_k}}}, \overline{n_{\iota_{q_k}}})$ is added in $\mathcal{N}$. The arcs of $\mathcal{N}$ are weighted. The weight of the arc $(\overline{n_{\iota_{j_k}}}, \overline{n_{\iota_{q_k}}})$ is obtained by applying the operator J^* on the topic sets $topicSet_{j_k}$ and $topicSet_{q_k}$ of ι_{j_k} and ι_{q_k}, respectively. Therefore, the weight of an arc in $\mathcal{N}$ belongs to the real interval $[0, 1]$; the higher this weight the higher the semantic similarity between the topics of the profiles $\overline{\mathcal{P}_{j_k}}$ and $\overline{\mathcal{P}_{q_k}}$ of ι_{j_k} and ι_{q_k}, respectively.
- For each c-arc in $\mathcal{M}$, which relates two instances $n_{\iota_{j_k}}$ and $n_{\iota_{j_q}}$ of the same object o_j in two different IoTs $\mathcal{I}_k$ and $\mathcal{I}_q$, the two nodes $\overline{n_{\iota_{j_k}}}$ and $\overline{n_{\iota_{j_q}}}$ in $\mathcal{N}$, corresponding to the nodes $n_{\iota_{j_k}}$ and $n_{\iota_{j_q}}$ in $\mathcal{M}$, are merged into a unique node $\overline{n_{\iota_j}}$. This node inherits all the arcs of $\overline{n_{\iota_{j_k}}}$ and $\overline{n_{\iota_{j_q}}}$.

At the end of these steps, it could happen that two or more arcs relate the same nodes $\overline{n}$ and $\overline{n'}$ in $\mathcal{N}$. In this case, all these arcs must be merged into a single arc. Clearly, it is necessary to determine the weight of this arc. Here, it appears reasonable that it must be higher than or equal to the maximum weight of the merged arcs. To reach this objective, our approach operates as follows. Let $\{(\overline{n}, \overline{n'}, \overline{w^1}), (\overline{n}, \overline{n'}, \overline{w^2}), \cdots, (\overline{n}, \overline{n'}, \overline{w^s})\}$ be the arcs to merge, ordered by decreasing weight. The new arc $(\overline{n}, \overline{n'}, \overline{w})$ will have a weight equal to:

$$\overline{w} = min\left(1, \overline{w_1} + \alpha \sum_{k=2..s} \overline{w_k}\right) \tag{10.29}$$

In other words, in the computation of $\overline{w}$, the arcs with the maximum weight will contribute with all their weight. All the other arcs will contribute to a lesser extent, with a fraction of their weight. This last is determined by means of the coefficient α.

Once the construction of $\mathcal{N}$ has been completed, the thematic views are derived by applying on $\mathcal{N}$ a graph clustering algorithm among the ones already existing in the literature (see [399] for a survey on them).

10.5.3 Discussion

An important issue about the supervised and the unsupervised approaches to address regards their scalability or, better, the possibility to use them in MIoTs comprising thousands or even millions of nodes.

With regard to this issue, first of all we observe that both approaches aim at deriving virtual IoTs which are, then, exploited by users to perform their desired tasks (such as querying). As a consequence, we can distinguish two moments in the life of a MIoT, namely: *(i)* the construction of virtual IoTs, which can be performed *offline*, and *(ii)* their usage, which is generally carried out *online*.

The first moment is computationally expensive because it involves several network operations in the supervised approach and a clustering activity in the unsupervised one. Clustering's computational cost is intrinsically exponential even if all the corresponding methods adopted in the reality are heuristic and most of them have a linear or a quadratic computational complexity. In any case, as pointed above, this task is performed offline and rarely because it is necessary only when many changes have been made in the MIoT.

The second moment is certainly less expensive; its cost depends on the size of the involved clusters; in fact, each user activity generally involves one or a few clusters. Concerning this aspect, it is important to verify: *(i)* if clustering is possible in presence of huge MIoTs, and *(ii)* how the size of clusters increases against the growth of the MIoT. As for the first point, we observe that, in the past, several algorithms have been specifically conceived to cluster a huge amount of elements [146]. Concerning the second point, instead, first we observe that the size of clusters can be determined by suitably tuning the parameters of the selected clustering algorithm. However, it could be interesting to verify how much the size of clusters increases if we maintain constant all the clustering algorithm parameters and the MIoT size increases. We decided to perform this experiment. It is described in detail in Section 10.6.6. Here, we evidence the obtained results, i.e., that when the MIoT size highly increases, the cluster size slightly grows, whereas the number of clusters increases very much. This is a positive result for our purposes because the parameter to monitor for investigating the performance obtained during the second moment is just cluster size.

Another important issue to investigate regards the possible existence of a unique framework handling all the objects of the MIoT and, therefore, in principle, thousands or millions of objects. With regard to this aspect, we evidence that, in the past, several attempts have been successfully performed in this direction (think, for instance, of the SIoT framework proposed in [39, 42]). Clearly, we understand that, in the future, the number of objects possibly belonging to a MIoT is enormously higher than the number of objects available in the past IoT frameworks. However, we point out that:

(i) our approach needs to store only the metadata of the involved objects, and these are small; *(ii)* the real objects can operate in a distributed environment thanks to the new available technologies, such as cloud, edge and fog computing, which can ease the organization and the management of distributed contexts.

10.6 Experiments

In this section, we present the experimental campaign that we carried out to evaluate the performance of our approach from several viewpoints. Specifically, we describe our dataset in a subsection, whereas, in the next ones, we illustrate our tests, along with the underlying motivations and the obtained results.

10.6.1 Adopted Dataset

To perform our experiments, we had the necessity to create several MIoTs with different sizes, ranging from hundreds to thousands of nodes. Since, currently, real MIoTs with the size and the variety handled by our model do not exist yet, we had to realize a MIoT simulator, i.e., a tool that, starting from real data, is capable of simulating MIoTs with certain characteristics specified by the user.

The MIoTs created by our simulator follow the model described in Section 10.3. In order to perform its task, our simulator carries out the following steps: *(i)* creation of objects; *(ii)* creation of object instances; *(iii)* creation of instance connections; *(iv)* creation of instance profiles.

Our MIoT simulator is also provided with a suitable interface allowing a user to "personalize" the MIoT to construct by specifying the desired values for several parameters, such as the number of nodes, the maximum number of instances of an object, and so forth.

To make "concrete" and "plausible" the created MIoT, our simulator leverages a real dataset. It regards the taxi routes in the city of Porto from July 1^{st} 2013 to June 30^{th} 2014. It can be found at the address `http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html`. Each route contains several Points of Interests corresponding to the GPS coordinates of the vehicle.

We partitioned the city of Porto in six areas and associated a real IoT with each of them. Our simulator associates an object with a given route recorded in the dataset and an object instance for each partition of a route belonging to an area. It creates a MIoT node for each instance and a c-arc for each pair of instances belonging to the same route. Furthermore, it creates an i-arc between two nodes of the same IoT if the length of the time interval between the corresponding routes is less than a certain threshold th_t. The weight of the i-arc indicates the length of this time interval. The

value of th_t can be specified through the constructor interface. Clearly, the higher th_t the more connected the constructed MIoT.

As far as instance profiles are concerned, since there are no thing profiles available (indeed, the concept of thing profile is one of the main novelties introduced in this chapter), we had to simulate them. However, we aimed to make them as real as possible. In order to increase the likelihood of constructed MIoTs, we performed a sentiment analysis task for each of the six areas in which we partitioned the city of Porto and for each day which the dataset refers to. For this purpose, we leveraged IBM Watson on the social media and blogs it uses as default. Having this data at disposal, our simulator assigns to each instance the most common topics (along with the corresponding occurrences) discussed in that area in the day on which the corresponding route took place. The constructed MIoTs are returned in a format that can be directly processed by the cypher-shell of Neo4J (see below).

Some features of the constructed MIoTs are reported in Table 10.1. The interested reader can find the MIoTs adopted in the experiments described in this section at the address `http://daisy.dii.univpm.it/miot/datasets/virtualIoTs`.

MIoT (size)	Number of arcs	Mean in-degree	Mean out-degree	Number of i-arcs	Number of c-arcs
$\mathcal{M}_1$ (176)	1176	6.29	6.61	980	126
$\mathcal{M}_2$ (301)	2050	7.76	7.74	1709	341
$\mathcal{M}_3$ (485)	3756	8.80	8.54	3130	626
$\mathcal{M}_4$ (778)	5866	8.89	9.11	4895	971
$\mathcal{M}_5$ (946)	7624	8.64	8.84	6422	1202
$\mathcal{M}_6$ (1256)	9860	7.87	7.98	7917	1943
$\mathcal{M}_7$ (1725)	12263	7.94	8.18	9964	2299
$\mathcal{M}_8$ (2028)	15568	8.22	8.38	12857	2711
$\mathcal{M}_9$ (3544)	26428	8.36	8.42	22718	3710
$\mathcal{M}_{10}$ (5024)	38642	8.44	8.54	33724	4918

Table 10.1. Main features of the constructed MIoTs

We carried out all the tests presented in this section on a server equipped with an Intel I7 Quad Core 7700 HQ processor and 16 GB of RAM with Ubuntu 16.04 operating system.

To implement our approaches we adopted:

- Python, powered with the NetworkX library, as programming language;
- Neo4J (Version 3.4.5) as underlying DBMS; we also exploited some plugins of Neo4J to perform community detection and to compute clustering coefficients.

10.6.2 Cohesion of the obtained topic-guided virtual IoTs

Our first test started from the idea that if our approach aims at extracting virtual thematic IoTs, they should present both a structural and a semantic cohesion higher than the corresponding ones characterizing the original IoTs of the MIoT. This experiment

was devoted to evaluate if this assumption is verified. We considered two well known structural cohesion parameters used in network analysis literature, namely *clustering coefficient* and *density* [439]. Both of them range in the real interval $[0, 1]$; the higher their value the higher the corresponding network cohesion. In the following, first we test the supervised approach and, then, we consider the unsupervised one.

Supervised approach

In this test, we run our supervised approach on ten MIoTs, $\mathcal{M}_1, \ldots, \mathcal{M}_{10}$, consisting of 176, 301, 485, 778, 946, 1256, 1725, 2028, 3544 and 5024 nodes. Clearly, the number of IoTs for each MIoT was equal to six, one for each area of the city of Porto that we have defined. For each MIoT, we submitted a set of 10 queries consisting of 1 (resp., 2, 4, 6, 8 and 10) word(s).

Each query returned a virtual thematic IoT for which we computed the corresponding clustering coefficient and density. Finally, we averaged the obtained results for each MIoT and for each set of queries, and we compared them with the average clustering coefficient and the average density of the corresponding real IoTs. The obtained results are reported in Tables 10.2 and 10.3.

MIoT (size)	Avg. clustering coeff. (real IoTs)	Avg. clustering coeff. (virtual IoTs)					
		$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.230	0.318	0.368	0.389	0.394	0.401	0.408
$\mathcal{M}_2$ (301)	0.272	0.343	0.388	0.419	0.424	0.434	0.446
$\mathcal{M}_3$ (485)	0.293	0.396	0.437	0.477	0.482	0.488	0.497
$\mathcal{M}_4$ (778)	0.353	0.447	0.478	0.503	0.508	0.511	0.517
$\mathcal{M}_5$ (946)	0.371	0.452	0.492	0.512	0.522	0.524	0.526
$\mathcal{M}_6$ (1256)	0.385	0.486	0.511	0.529	0.530	0.532	0.535
$\mathcal{M}_7$ (1725)	0.386	0.501	0.524	0.536	0.537	0.538	0.539
$\mathcal{M}_8$ (2028)	0.388	0.519	0.536	0.541	0.541	0.542	0.543
$\mathcal{M}_9$ (3544)	0.392	0.522	0.540	0.544	0.544	0.545	0.546
$\mathcal{M}_{10}$ (5024)	0.395	0.534	0.546	0.546	0.546	0.547	0.548

Table 10.2. Values of the clustering coefficient for real and virtual IoTs against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average density (real IoTs)	Average density (virtual IoTs)					
		$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.348	0.260	0.264	0.280	0.289	0.296	0.301
$\mathcal{M}_2$ (301)	0.262	0.292	0.303	0.309	0.315	0.320	0.324
$\mathcal{M}_3$ (485)	0.274	0.390	0.395	0.400	0.402	0.405	0.408
$\mathcal{M}_4$ (778)	0.269	0.476	0.483	0.490	0.501	0.509	0.514
$\mathcal{M}_5$ (946)	0.276	0.492	0.509	0.521	0.536	0.534	0.556
$\mathcal{M}_6$ (1256)	0.284	0.547	0.556	0.567	0.572	0.576	0.581
$\mathcal{M}_7$ (1725)	0.278	0.582	0.582	0.594	0.598	0.598	0.601
$\mathcal{M}_8$ (2028)	0.273	0.609	0.610	0.620	0.626	0.630	0.639
$\mathcal{M}_9$ (3544)	0.269	0.626	0.628	0.630	0.634	0.636	0.637
$\mathcal{M}_{10}$ (5024)	0.262	0.636	0.636	0.638	0.638	0.640	0.642

Table 10.3. Values of the density for real and virtual IoTs against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

From the analysis of these tables, we can observe that, in almost all circumstances, the values of both clustering coefficient and density are higher or much higher for the virtual thematic IoTs than for the real ones. This is clearly a confirmation of the goodness of our supervised approach, which returns topic-guided IoTs more cohesive than the original ones. We also observe that when $|Q|$ increases, the values of both clustering coefficient and density increases. This can be explained by observing that, in processing Q, our approach takes the portions of networks containing at least one keyword of Q. When $|Q|$ increases, the portion of networks selected by our approach increases too, and the probability of selecting a very high number of edges (i.e., a number so high to lead to an increase of clustering coefficient and density) increases as well.

Unsupervised approach

In this test, we run our unsupervised approach, powered with the Louvain graph clustering algorithm [68] as underlying engine, on the same MIoTs described in Section 10.6.2. For each MIoT, we computed the average clustering coefficient and the average density of real and virtual IoTs. The obtained results are reported in Table 10.4.

MIoT (size)	Average clustering coefficient		Average density	
	Real IoTs	Virtual IoTs	Real IoTs	Virtual IoTs
$\mathcal{M}_1$ (176)	0.230	0.473	0.348	0.315
$\mathcal{M}_2$ (301)	0.272	0.499	0.262	0.350
$\mathcal{M}_3$ (485)	0.293	0.500	0.274	0.375
$\mathcal{M}_4$ (778)	0.353	0.511	0.269	0.318
$\mathcal{M}_5$ (946)	0.372	0.509	0.276	0.316
$\mathcal{M}_6$ (1256)	0.385	0.506	0.284	0.314
$\mathcal{M}_7$ (1725)	0.386	0.522	0.280	0.328
$\mathcal{M}_8$ (2028)	0.388	0.535	0.273	0.360
$\mathcal{M}_9$ (3544)	0.394	0.547	0.271	0.364
$\mathcal{M}_{10}$ (5024)	0.398	0.562	0.269	0.368

Table 10.4. Values of both clustering coefficient and density of real and virtual IoTs against the size of MIoTs (unsupervised approach)

From the analysis of this table we can observe that, in this case, analogously to what happened for the supervised approach, the cohesion level of the virtual IoTs is higher or much higher than the corresponding ones of the real original IoTs. Interestingly, both clustering coefficient and density values obtained by the unsupervised approach are generally higher than those returned by the supervised one, at least when the MIoT size is small. Instead, when the MIoT size is large, they become lower than the ones of the supervised approach. Actually, the increase of both clustering coefficient and density when the MIoT size increases is significant for the supervised approach, whereas it is more limited for the unsupervised one.

10.6.3 Average fraction of merged c-nodes and analysis of node distribution in virtual IoTs

Another quality parameter for virtual IoTs returned by our approach regards the average number of merged c-nodes present in each of them. Indeed, the presence of merged c-nodes in an IoT is an indicator of the fact that this IoT is capable of connecting concepts coming from different real IoTs, and, therefore, from concepts whose relationships would have been uncaptured otherwise, or, in other words, that the knowledge it is presenting is new and did not exist previously. Clearly, the higher the fraction of merged c-nodes and the higher the fraction of different original IoTs they belong to, the higher the connecting capability of virtual IoTs.

Also for this experiment, we considered the ten MIoTs described in Section 10.6.2 and performed the same tasks illustrated therein for both the supervised and the unsupervised approaches. The obtained results are reported in Tables 10.5, 10.6 and 10.7.

MIoT (size)	Average fraction of merged c-nodes					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.304	0.455	0.517	0.532	0.554	0.572
$\mathcal{M}_2$ (301)	0.380	0.515	0.608	0.627	0.652	0.679
$\mathcal{M}_3$ (485)	0.539	0.661	0.782	0.798	0.813	0.823
$\mathcal{M}_4$ (778)	0.690	0.786	0.860	0.874	0.883	0.892
$\mathcal{M}_5$ (946)	0.724	0.812	0.884	0.898	0.916	0.924
$\mathcal{M}_6$ (1256)	0.808	0.883	0.939	0.943	0.946	0.948
$\mathcal{M}_7$ (1725)	0.862	0.908	0.952	0.961	0.961	0.963
$\mathcal{M}_8$ (2028)	0.908	0.959	0.974	0.975	0.976	0.977
$\mathcal{M}_9$ (3544)	0.928	0.963	0.976	0.977	0.977	0.978
$\mathcal{M}_{10}$ (5024)	0.936	0.968	0.978	0.979	0.980	0.981

Table 10.5. Average fraction of merged c-nodes against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average fraction of involved real IoTs					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.373	0.467	0.488	0.476	0.452	0.448
$\mathcal{M}_2$ (301)	0.365	0.469	0.525	0.501	0.488	0.480
$\mathcal{M}_3$ (485)	0.482	0.477	0.448	0.442	0.435	0.432
$\mathcal{M}_4$ (778)	0.457	0.432	0.418	0.415	0.413	0.411
$\mathcal{M}_5$ (946)	0.455	0.482	0.624	0.628	0.647	0.644
$\mathcal{M}_6$ (1256)	0.453	0.514	0.805	0.864	0.917	0.924
$\mathcal{M}_7$ (1725)	0.482	0.577	0.815	0.872	0.917	0.924
$\mathcal{M}_8$ (2028)	0.514	0.672	0.833	0.898	0.917	0.924
$\mathcal{M}_9$ (3544)	0.584	0.704	0.844	0.905	0.924	0.926
$\mathcal{M}_{10}$ (5024)	0.624	0.727	0.888	0.911	0.928	0.934

Table 10.6. Average fraction of real IoTs involved in a virtual IoT against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

From the analysis of these tables, we observe that both the supervised and the unsupervised approaches return satisfying results. As for the supervised approach,

MIoT (size)	Average fraction of merged c-nodes	Average fraction of involved real IoTs
$\mathcal{M}_1$ (176)	0.227	0.361
$\mathcal{M}_2$ (301)	0.306	0.353
$\mathcal{M}_3$ (485)	0.309	0.357
$\mathcal{M}_4$ (778)	0.342	0.356
$\mathcal{M}_5$ (946)	0.334	0.359
$\mathcal{M}_6$ (1256)	0.326	0.361
$\mathcal{M}_7$ (778)	0.332	0.360
$\mathcal{M}_8$ (2028)	0.335	0.358
$\mathcal{M}_9$ (3544)	0.341	0.371
$\mathcal{M}_{10}$ (5024)	0.344	0.378

Table 10.7. Average fraction of merged c-nodes and average fraction of real IoTs involved in a virtual IoT against the size of MIoTs (unsupervised approach)

we can observe that the fraction of merged c-nodes increases when the size of MIoT increases. Furthermore, we can also observe a slight increase of this fraction when $|Q|$ increases. The same trends can be observed for the average fraction of involved real IoTs, even if, for this parameter, its increase against the increase of $|Q|$ is more pronounced. As for the unsupervised approach, we can observe that the average fraction of merged nodes is always very high, independently of the MIoT size. By contrast, in this case, the fraction of involved real IoTs is quite high even if lower than the ones generally observed for the supervised approach. Furthermore, its value does not significantly change when the MIoT size increases.

In order to deepen this investigation, for each virtual IoT, we compared the distribution of its nodes against the real IoTs they belong to. Indeed, if almost all the nodes of a virtual IoT derive from only one real IoT, the information contribution provided by the virtual IoT would be very small because it would be analogous to the one provided by the corresponding real IoT. By contrast, if the nodes of a virtual IoT homogeneously derive from several real IoTs, then the knowledge it provides is really new, and this knowledge would be uncaptured and lost if the new IoT had not been extracted. On the basis of this reasoning, we evaluated the heterogeneity of the provenance of the various nodes of each virtual IoT (see below). For this purpose, we adapted the Herfindahl Index [203] to our context. This index is very used in several research fields of Economics from several decades; for instance, it is exploited to evaluate the concentration degree in an industry.

In order to adapt the Herfindahl Index to our scenario, consider a MIoT $\mathcal{M}$ consisting of s real IoTs $(\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_s)$. Consider, also, a virtual IoT $\mathcal{V}_j$ derived by either the supervised or the unsupervised approach. Let n_j be the number of nodes of $\mathcal{V}_j$ and let $\frac{n_{j_k}}{n_j}$, $1 \leq k \leq s$, be the fraction of the nodes of $\mathcal{V}_j$ belonging to $\mathcal{R}_k$ (i.e., the k^{th} real IoT of the MIoT). The Herfindahl Index H_j of $\mathcal{V}_j$ is defined as $\sum_{k=1}^{s} \left(\frac{n_{j_k}}{n_j}\right)^2$. H_j ranges in the real interval $\left[\frac{1}{s}, 1\right]$; the higher its value, the higher the concentration degree of the nodes of $\mathcal{R}_k$ in $\mathcal{V}_j$. Clearly, as previously pointed out, one property desired for our approach is the ability to construct virtual IoTs connecting nodes that

belong to different real IoTs in such a way as to extract knowledge that would be lost otherwise. If we report this property to the Herfindahl Index, this implies to obtain a value of this index as lower as possible[3].

We computed the average Herfindahl Index of the thematic IoTs returned by both the supervised and the unsupervised approaches by considering the ten MIoTs described in Section 10.6.2 and performing the same tasks illustrated therein. The obtained results are reported in Tables 10.8 and 10.9.

MIoT (size)	Average Herfindhal Index					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.207	0.186	0.177	0.175	0.173	0.172
$\mathcal{M}_2$ (301)	0.204	0.183	0.174	0.173	0.172	0.171
$\mathcal{M}_3$ (485)	0.178	0.173	0.170	0.170	0.169	0.168
$\mathcal{M}_4$ (778)	0.172	0.172	0.170	0.170	0.169	0.168
$\mathcal{M}_5$ (946)	0.172	0.170	0.169	0.169	0.169	0.168
$\mathcal{M}_6$ (1256)	0.173	0.168	0.167	0.169	0.168	0.167
$\mathcal{M}_7$ (1725)	0.170	0.168	0.167	0.169	0.168	0.167
$\mathcal{M}_8$ (2028)	0.168	0.167	0.167	0.167	0.167	0.167
$\mathcal{M}_9$ (3544)	0.168	0.167	0.167	0.167	0.167	0.167
$\mathcal{M}_{10}$ (5024)	0.167	0.167	0.167	0.167	0.167	0.167

Table 10.8. Average Herfindahl Index of virtual IoTs against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average Herfindahl Index
$\mathcal{M}_1$ (176)	0.658
$\mathcal{M}_2$ (301)	0.543
$\mathcal{M}_3$ (485)	0.658
$\mathcal{M}_4$ (778)	0.636
$\mathcal{M}_5$ (946)	0.654
$\mathcal{M}_6$ (1256)	0.694
$\mathcal{M}_7$ (1725)	0.656
$\mathcal{M}_8$ (2028)	0.635
$\mathcal{M}_9$ (3544)	0.664
$\mathcal{M}_{10}$ (5024)	0.686

Table 10.9. Average Herfindahl Index of virtual IoTs against the size of MIoTs (unsupervised approach)

These tables evidence that also the analysis based on object distribution and Herfindahl Index returns very satisfying results that confirm and strengthen those obtained by examining the average fraction of merged nodes involved in a virtual IoT. Interestingly, as for this parameter, we observe that the supervised approach returns excellent results, very close to the best ones. By contrast, the unsupervised approach returns good results, even if those returned by the supervised approach are better.

[3] Consider that, since we have six real IoTs in our MIoTs, the minimum value of the Herfindahl Index is $\frac{1}{6} = 0.167$.

10.6.4 Computation time

In this experiment, we aimed at evaluating the variation of the computation time of both the supervised and the unsupervised approaches against the variation of the size of the involved MIoT. Furthermore, as for the supervised approach, we also evaluated the variation of the computation time against the variation of the size of queries.

To perform this task, we considered the ten MIoTs described in Section 10.6.2 and carried out the same tasks illustrated therein. Finally, we measured the corresponding average computation times. The obtained results are reported in Figures 10.1, 10.2 and 10.3.

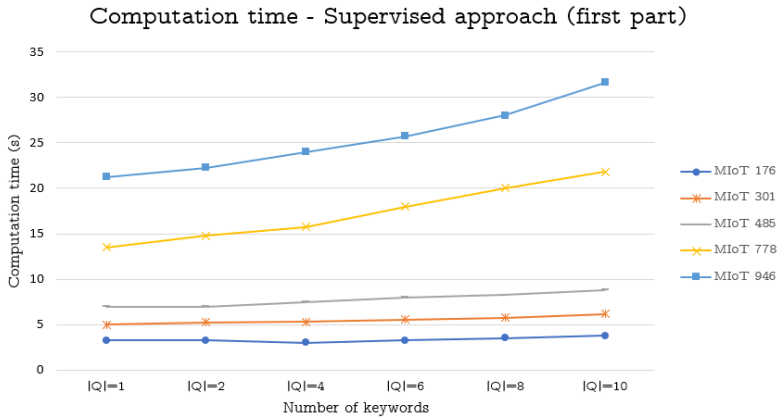


Fig. 10.1. Computation time (in seconds) against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach) - first part

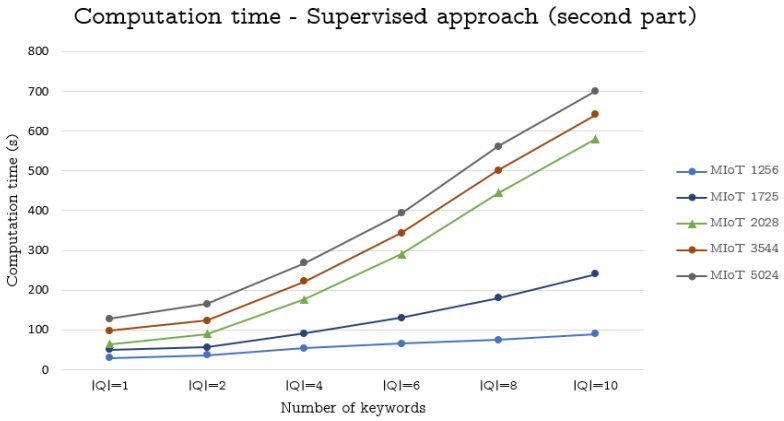


Fig. 10.2. Computation time (in seconds) against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach) - second part

From the analysis of these figures, we can observe that our approaches obtain satisfying results. Specifically, as for the supervised approach, the computation time

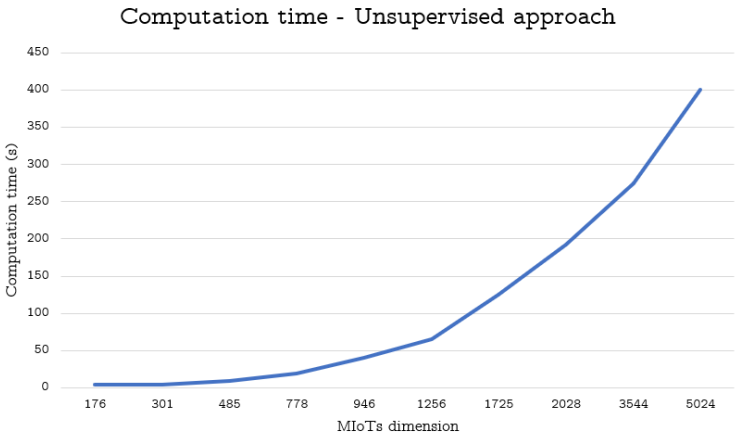


Fig. 10.3. Computation time (in seconds) against the size of MIoTs (unsupervised approach)

is always very low for MIoTs having at most 1256 nodes. Instead, for MIoTs with more than 2028 nodes, the computation time is low for $|Q| = 1$ or $|Q| = 2$. Then, it increases, even if it remains acceptable for $|Q| = 4$ and $|Q| = 6$, whereas it becomes excessive for $|Q| = 8$ and $|Q| = 10$. However, with regard to this fact, we must point out that queries consisting of 8 or 10 keywords are very uncommon[4].

As for the unsupervised approach, its computation time is still acceptable also for 2028 nodes. It starts to become excessive with MIoTs consisting of at least 10000 nodes.

10.6.5 Our approaches' capability of improving the efficiency of information dissemination

This experiment was devoted to measure the efficiency of both supervised and unsupervised approaches. The rationale underlying this experiment is that if some information must be transferred from a source object o_s to a target one o_t, the number of objects to be contacted for this task should be minimized. At the same time, if an object is involved in an information dissemination task, it would be desiderable that the information it is transmitting is also useful for it (which, in our case, means that it is in line with the interests of its profile).

In order to perform this experiment, we randomly selected some pairs of (source, target) nodes from our MIoT. Let (n_s, n_t) be one of these pairs. We verified if there

[4] It is worth pointing out that the topics considered by our approach for constructing a thing's profile are extremely generic and heterogeneous. As a consequence, in our scenario, a query with 8 or 10 keywords would encompass a great number of different topics and, as such, it would not be generally able to capture a clear and specific desire of a user.

existed at least one virtual IoT comprising both n_s and n_t[5]. In the negative case, we discarded that pair. Let $\mathcal{V}$ be a virtual IoT comprising both n_s and n_t.

After this, we computed the number $num_{st}^{\mathcal{V}}$ (resp., $\widehat{num_{st}^{\mathcal{V}}}$) of MIoT nodes involved in the dissemination of information in presence (resp., absence) of the virtual IoT $\mathcal{V}$. Specifically, we computed $num_{st}^{\mathcal{V}}$ by performing the information dissemination task only through its nodes; instead, we obtained $\widehat{num_{st}^{\mathcal{V}}}$ by performing the same task on the whole MIoT. Finally, we computed: $f_{st} = \frac{num_{st}^{\mathcal{V}}}{\widehat{num_{st}^{\mathcal{V}}}}$. Clearly, the lower f_{st}, the higher the contribution of the virtual IoTs in reducing the number of nodes necessary for the information dissemination task and, consequently, the higher the contribution that our virtual IoT detection approach can provide to information dissemination.

We computed the average values of f_{st} by operating on the ten MIoTs introduced in Section 10.6.2 and by performing the same tasks described therein for both the supervised and the unsupervised approaches. The obtained results are reported in Tables 10.10 and 10.11.

MIoT (size)	Average f_{st}					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.144	0.220	0.290	0.304	0.336	0.347
$\mathcal{M}_2$ (301)	0.126	0.170	0.177	0.175	0.178	0.179
$\mathcal{M}_3$ (485)	0.104	0.112	0.074	0.052	0.041	0.037
$\mathcal{M}_4$ (778)	0.057	0.051	0.028	0.038	0.047	0.049
$\mathcal{M}_5$ (946)	0.048	0.034	0.022	0.028	0.032	0.024
$\mathcal{M}_6$ (1256)	0.031	0.015	0.017	0.011	0.007	0.007
$\mathcal{M}_7$ (1725)	0.026	0.014	0.011	0.010	0.008	0.008
$\mathcal{M}_8$ (2028)	0.016	0.010	0.009	0.009	0.009	0.009
$\mathcal{M}_9$ (3544)	0.012	0.009	0.009	0.009	0.009	0.009
$\mathcal{M}_{10}$ (5024)	0.011	0.008	0.007	0.007	0.007	0.007

Table 10.10. Average values of f_{st} against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average f_{st}
$\mathcal{M}_1$(176)	0.904
$\mathcal{M}_2$(301)	0.722
$\mathcal{M}_3$(485)	0.635
$\mathcal{M}_4$(778)	0.584
$\mathcal{M}_5$(946)	0.580
$\mathcal{M}_6$(1256)	0.576
$\mathcal{M}_7$(1725)	0.516
$\mathcal{M}_8$(2028)	0.477
$\mathcal{M}_9$(3544)	0.452
$\mathcal{M}_{10}$(5024)	0.426

Table 10.11. Average values of f_{st} against the size of MIoTs (unsupervised approach)

From the analysis of these tables we can observe that both the supervised and the unsupervised approaches really contribute to decrease the number of the nodes

[5] This is always true for the unsupervised approach, whereas it could not happen for the supervised one.

of a MIoT involved in the information dissemination, and, therefore, to increase the efficiency of this task. As for the supervised approach, we observe that the decrease of the number of involved nodes is always high. It becomes very high as the MIoT size and the number of keywords composing the query increase. As for the unsupervised approach, we observe that it leads to a decrease of the number of the MIoT nodes involved in the dissemination task. However, this decrease is minimum for small MIoTs, whereas it becomes significant for large ones (i.e., for MIoTs with a number of nodes higher than 1256).

We performed a second experiment in this direction. Specifically, given a pair (n_s, n_t) of a MIoT such that information must be disseminated from n_s to n_t and there exists at least one virtual IoT $\mathcal{V}$ comprising both n_s and n_t, we computed the fraction $g_{st}^{\mathcal{V}}$ (resp., $\widehat{g_{st}^{\mathcal{V}}}$) of the nodes of the MIoT involved in the diffusion of information from n_s to n_t and having at least one content of the disseminated information registered in their profile (which implies that, in principle, they could benefit from the information they are required to disseminate). As in the previous experiment, we computed $g_{st}^{\mathcal{V}}$ by assuming the existence of $\mathcal{V}$ and, hence, by performing the information dissemination task through it; by contrast, we computed $\widehat{g_{st}^{\mathcal{V}}}$ by carrying out the information dissemination task through the whole MIoT. Finally, we computed $g_{st} = \frac{g_{st}^{\mathcal{V}}}{\widehat{g_{st}^{\mathcal{V}}}}$. Roughly speaking, it denotes how much the presence of the virtual IoT $\mathcal{V}$ can contribute to require information dissemination tasks only to nodes possibly benefiting of it. A value of this coefficient higher than 1 denotes a positive contribution of $\mathcal{V}$; the higher this value the higher the contribution. As in the previous experiment, we computed the average values of g_{st} by operating on the ten MIoTs introduced in Section 10.6.2 and by performing the same tasks described therein for both the supervised and the unsupervised approaches. The obtained results are reported in Tables 10.12 and 10.13.

MIoT (size)	Average g_{st}					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	4.018	2.792	2.223	1.918	1.331	1.321
$\mathcal{M}_2$ (301)	3.563	2.619	2.445	2.009	1.683	1.664
$\mathcal{M}_3$ (485)	3.269	2.370	1.426	1.528	1.626	1.674
$\mathcal{M}_4$ (778)	3.130	2.168	2.367	1.916	1.494	1.325
$\mathcal{M}_5$ (946)	3.232	2.102	1.864	1.712	1.461	1.391
$\mathcal{M}_6$ (1256)	3.467	1.979	1.378	1.412	1.438	1.452
$\mathcal{M}_7$ (1725)	3.476	2.224	1.414	1.444	1.494	1.492
$\mathcal{M}_8$ (2028)	3.496	2.669	1.489	1.491	1.521	1.545
$\mathcal{M}_9$ (3544)	3.507	2.712	1.612	1.624	1.631	1.632
$\mathcal{M}_{10}$ (5024)	3.517	2.926	1.783	1.841	1.864	1.874

Table 10.12. Average values of g_{st} against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average g_{st}
$\mathcal{M}_1$ (176)	1.341
$\mathcal{M}_2$ (301)	1.269
$\mathcal{M}_3$ (485)	1.211
$\mathcal{M}_4$ (778)	1.177
$\mathcal{M}_5$ (946)	1.173
$\mathcal{M}_6$ (1256)	1.171
$\mathcal{M}_7$ (1725)	1.194
$\mathcal{M}_8$ (2028)	1.273
$\mathcal{M}_9$ (3544)	1.281
$\mathcal{M}_{10}$ (5024)	1.301

Table 10.13. Average values of g_{st} against the size of MIoTs (unsupervised approach)

The analysis of these tables is a further confirmation of the efficiency of our approach. Indeed, thanks to the presence of virtual IoTs, the fraction of nodes participating to the spreading of information that can also benefit from this task increases remarkably.

The results of Tables 10.10 and 10.11, along with the ones of Tables 10.12 and 10.13, agree to evidence that the discovery of virtual IoTs is highly beneficial in terms of efficiency for the information dissemination task in a MIoT. In this case, the contribution of $\mathcal{V}$ in increasing the efficiency of the spreading task, by limiting it mainly to nodes that could benefit from the information they are disseminating, is very high for the supervised approach when $|Q| = 1$ or $|Q| = 2$. When $|Q|$ increases, this contribution decreases, even if it remains still significant. As for the unsupervised approach, the contribution of $\mathcal{V}$ can be always observed even if it is less evident than the one characterizing the supervised approach.

10.6.6 Number and size of returned virtual IoTs

This last experiment makes sense only for the unsupervised approach. Through it we aimed at investigating how the number and the size of returned virtual IoTs (and, therefore, the number and the size of returned clusters) vary when the MIoT size increases. To make this experiment significant, we maintained constant all the parameters of the adopted clustering algorithm. We considered the MIoTs $\mathcal{M}_1 \cdots \mathcal{M}_{10}$ used in the previous experiments because, in this way, we had the possibility to investigate MIoT sizes ranging from 176 to 5024 nodes. We report the obtained results in Table 10.14.

From the analysis of this table we can observe that the average size of virtual IoTs:

- increases when the MIoT size ranges from 176 to 946;
- slightly increases when the MIoT size ranges from 946 to 2028;
- remains essentially constant when the MIoT size is higher than 2028.

In the meantime, the number of clusters:

MIoT (size)	Average size of virtual IoTs	Number of virtual IoTs
$\mathcal{M}_1$ (176)	22.44	10
$\mathcal{M}_2$ (301)	28.21	13
$\mathcal{M}_3$ (485)	36.64	16
$\mathcal{M}_4$ (778)	40.82	22
$\mathcal{M}_5$ (946)	44.66	24
$\mathcal{M}_6$ (1256)	46.74	30
$\mathcal{M}_7$ (1725)	48.12	39
$\mathcal{M}_8$ (2028)	50.24	45
$\mathcal{M}_9$ (3544)	50.46	78
$\mathcal{M}_{10}$ (5024)	50.64	105

Table 10.14. Average size and number of virtual IoTs against the increase of the MIoT size (unsupervised approach)

- slightly increases when the MIoT size ranges from 176 to 946;
- increases when the MIoT size ranges from 946 to 2028;
- highly increases when the MIoT size is higher than 2028.

The obtained results are extremely interesting because they confirm the soundness of the reasoning made in Section 3.3.7. In particular, this experiment confirms the scalability of our approach. As a matter of fact, after the virtual IoTs have been constructed offline, their usage for querying and for the other tasks of interest for the user can be performed online. Now, we observed that the number of available virtual IoTs highly increases when the MIoT size increases. However, because the size of each virtual IoT is only slightly impacted by the growth of the corresponding MIoT, and because user tasks generally involve one or at most a few of available virtual IoTs, we can conclude that our approach is scalable with respect to the size variation of the MIoT.

Part IV

Innovation Management

In this part, we apply our network-based model and the associated social network-based approach to support decision making in the field of innovation management. This part is organized as follows: in Chapter 11, we propose a well-tailored centrality measure for evaluating patents and their citations. In Chapter 12, we present a new approach to extract knowledge patterns about research activities and hubs in a set of countries. Finally, in Chapter 13, we introduce new metrics specifically conceived to evaluate the innovation level of each country based on patent data.

11

Evaluating patents and their citations

11.1 Introduction

Patents have been largely investigated in the past scientific literature [8, 285, 446, 141, 424, 252]. In fact, their analysis can supply a large amount of information concerning both the state of art and the protagonists of a certain Research & Development (R&D) field [466, 156, 181, 200, 208, 308, 410, 281]. This also because the submission of a patent is usually the first public claim of a new invention or innovation. Patent analysis allows decision makers to investigate the experiences of other (possible competitor) institutions and/or countries, in such a way as to know the past and the current R&D activities in the fields of interest, to delineate their evolution and to foresee their future developments. Furthermore, patent analysis allows the construction of a detailed picture of the R&D cooperations among different institutions and/or countries and can be an indicator of geo-political evolutions happening all over the world [109, 405, 73].

Most of the past approaches for patent analysis were based on classical statistics. However, the impressive development of innovations in all the R&D fields is leading to a huge increase of patent data. Therefore, it is reasonable to foresee that, in the next future, Big Data centered techniques will be compulsory to fully exploit the potential of patent data. In this last scenario, the adoption of approaches based on network analysis is extremely promising [458, 106, 107, 258, 476, 87]. As a matter of facts, network analysis allows a full comprehension and a complete management of those phenomena where relationships among objects to investigate play the key role and, at the same time, the corresponding variables are strictly related to each other. This is exactly the future scenario characterizing patent and innovation management, and, at the same time, it is the "worst-case scenario" for classic statistic-based approaches, which present several limitations when operating therein [439].

As a confirmation of the adequacy of network analysis for patent investigation, in the past literature, several approaches to facing this issue can be found (see, for instance, [87, 214, 142, 209, 469]).

Centrality is one of the most investigated issues in network analysis. It aims at measuring the importance of a node in a network. It allows experts: *(i)* to measure the relevance and the criticity of nodes in their networks; *(ii)* to define forms of distance between network nodes or areas; *(iii)* to measure the cohesion degree of a subnetwork; *(iv)* to identify cohesive subnetworks or network communities.

In the past, several centrality measures have been proposed in the literature [94, 386, 162, 186, 161, 423, 80]. Among them, the most general and best known ones are: *(i)* degree centrality, based on the number of arcs incoming in, or outgoing from, each node; *(ii)* closeness centrality, based on distances between nodes; *(iii)* betweenness centrality, based on the shortest paths connecting pairs of nodes; *(iv)* eigenvector centrality, based on both the number and the centrality of nodes whose outgoing arcs are incident on the nodes of interest. All these measures, as well as the other ones proposed in the literature, could be adopted in the investigation of patents. However, they are not tailored to this scenario and could return approximate results. This because patents have a very relevant peculiarity that is not found elsewhere (for instance, in scientific papers [153]), in that, if a patent p_i cites a patent p_j, then p_i looses a part of its value.

If we report this reasoning to the network analysis context, we have that, for a node, having incoming arcs is extremely positive; by contrast, having outgoing arcs is negative. Past centrality measures certainly distinguish between these two kinds of arc; for instance, degree centrality distinguishes between indegree and outdegree [193]. However, they do not combine centrality values originated from the incoming arcs with those derived from the outgoing ones. What is missing is precisely a centrality measure that first assigns a positive ranking to incoming arcs and a negative ranking to outgoing ones and, then, combines these rankings to obtain a unique value.

In this paper, we aim at providing a contribution in this setting. In fact, we propose a well-tailored centrality measure for evaluating patents and their citations.

For this purpose, we preliminarily introduce a suitable support directed network, whose nodes represent patents. An arc from a node v_i to a node v_j indicates that the patent represented by v_i cited the patent represented by v_j.

After this, we introduce our centrality measures, namely "Naive Patent Degree" and "Refined Patent Degree", and we show that they are well tailored to capture the specificities of the patent scenario. To investigate the adequacy of our centrality measures, we carried out several experiments. The corresponding patent data derives from PATSTAT-ICRIOS database [108]. This is a large dataset about patents constructed

and maintained by the Invernizzi Centre of Research and Innovation, Organization and Strategy (ICRIOS) at Bocconi University. It stores patent data, from 1978 to the current year, coming from about 90 patent offices worldwide, including, of course, the most important and largest ones, such as European Patent Office (EPO) and United States Patent and Trademark Office (USPTO).

Finally, we present three possible applications of our measures, namely: *(i)* the computation of the "scope" of a patent, whose purpose is the evaluation of the width and the strength of the influence of a patent on a given R&D field; *(ii)* the computation of the lifecycle of a patent; *(iii)* the detection of the so-called "power patents", i.e., the most relevant patents, and the investigation of the importance, for a patent, to be cited by a power patent.

The plan of this paper is as follows: in Section 11.2, we present related literature. In Section 11.3, we illustrate the patent database that we used for our experiments, and the support network model that we defined to represent patents and their relationships. In Section 11.4, we present our centrality measures and evaluate them. In Section 11.5, we describe the three applications of our centrality measures that we mentioned above.

11.2 Related Work

Centrality has always been one of the core topics of network analysis and has been largely investigated in the literature. It allows people to quantify the importance of nodes in their network and to understand the structural properties of this last one. As a matter of facts, already [366] developed a self-consistent methodology for determining citation-based influence measures for scientific journals, subfields and fields. Specifically, these authors formulate an eigenvalue problem leading to a size-independent influence weight for each journal or aggregate. Then, they define two other measures, namely the influence per publication and the total influence. Finally, they present some hierarchical influence diagrams and numerical data to display inter-relationships for journals on physics. In the same years, [162] examined and explained the role of centrality metrics in network analysis.

As illustrated in detail in [284, 111], the influence of a node mainly depends on its position in the corresponding network, as well as on the structural properties of this last one. Centrality metrics aim at assigning a rank to each network node, summarizing its importance in the network. As previously pointed out, this rank is strictly related to the needs of the application scenario, which the network refers to. Since these needs can be heterogeneous, several different metrics have been proposed in the past network analysis literature.

The study of the neighborhood of a node is adopted as the starting point of some of the most important centrality metrics. In this context, degree centrality is one of the most famous metrics; it aims at measuring the visibility of a node within its network. Degree centrality presents several strengths but also some weaknesses. This is the reason why, in the literature, researchers proposed some approaches that try to overcome the problems of this metric. An example is ClusterRank, proposed in [94]; it also considers clustering coefficient in the score computation. In [134], the authors, starting from the observation that the position of a node is more important than its degree for measuring its relevance, apply k-core decomposition. It iteratively breaks down the network according to the residual degree of its nodes. K-core decomposition is considered as one of the most valid approaches to understanding the influence of a node and its role in information diffusion. Another well known centrality measure is h-index [201], which returns the influence of a user in a social network.

Another family of centrality approaches is based on the number of paths, which a node is involved in. In this path-based centrality, the higher the number of paths where a certain node is present the higher the node's importance. Closeness centrality [386], eccentricity centrality [186] and betweenness centrality [161] belong to this family of approaches. From a general point of view, a node with a high closeness centrality can have access to a high number of communications; therefore, it can perform a high control on information flow. Instead, a node with a high betweenness centrality, in most cases, operates as a bridge between two communities; therefore, it can have a strong control on information exchange. Other techniques belonging to this family of centrality metrics are Kats centrality [232], subgraph centrality [145], and information index [423].

As pointed out in [465], in most cases, centrality does not depend only on the number of neighbors of a node on the paths it is involved in. In some cases, not only the number of neighbors, but also their relevance is important to assess the relevance of a node in its network. Starting from this consideration, authors have defined a third family of centrality measures. Eigenvector centrality [70], PageRank [80] and HITs [239] are the most known metrics of this family.

Even if centrality is one of the most important topics in network analysis, it was rarely adopted for investigating the relevance of a patent based on citations. Actually, the idea of analyzing patents based on their citations was proposed by Seidel in 1949 [401]. From that time, a large variety of tools for performing this analysis has been proposed in the literature. Network analysis is one of the most adopted tools because it allows the creation of suitable networks representing patent citations.

Bibliometrics is certainly an optimum starting point for patent investigation, as it shares many common aspects with patent analysis. Clearly, besides many similarities,

paper and patent citations also present several significant differences, as evidenced in [307].

If we focus on patent citations, several variegate approaches to investigating patents based on them have been proposed in the past. For instance, the authors of [469] consider both direct and indirect citations, as well as patent couplings co-citations. An approach to investigating patent outliers is described in [381], whereas the small world phenomenon in the context of patent citation networks is analyzed in [214]. The definition of the lifecycle of a given technology starting from patent citation networks is proposed in [211], whereas the technological focus of patents is studied in [210].

In several cases, the typical problems of network analysis are investigated in the context of patent citation networks. For instance, the approach to analyzing network connectivity proposed in [212] is extended to patent citation networks in [54, 157, 443]. Specifically, [54] shows how the analysis of network connectivity can be extended to the patent scenario for detecting reliable knowledge on technological evolutions. [157] exploits network connectivity to reconstruct the most relevant technological trajectories of data communication standards. [443] performs a similar investigation but for fuel cells technology.

Finally, the application of the standard centrality metrics to patent citation networks has been proposed in very few cases. For instance, the authors of [87] propose an approach to determining the relevance of companies in the industry they operate on, based on the application of classic centrality metrics on the citation networks of the patents published by them. An analogous effort can be found in [92], but for Intelligent Transportation System companies. The authors of [257] apply degree centrality, betweenness centrality and closeness centrality on patent citation networks to investigate several mechanisms underlying technological innovations. Finally, in [144, 303], the authors carefully examine the usage of PageRank in patent citation networks, and evidence its strengths and weaknesses.

However, to the best of our knowledge, none of the approaches proposing the application of centrality measures to patent citation networks considers the main peculiarity of this scenario, i.e., that, if a patent p_i cites a patent p_j, then the value of p_i decreases. By contrast, this important feature represents the core of our approach.

11.3 Preliminaries

11.3.1 Patent Database

Data regarding patents adopted in our analyses has been taken from PATSTAT-ICRIOS database [108]. This is a large database about patents handled by ICRIOS Center at Bocconi University.

PATSTAT (i.e., EPO worldwide PATent STATistical database) is a database storing raw data about patents. It was constructed by EPO in cooperation with the World Intellectual Property Organization (WIPO), OECD and Eurostat. It is currently managed by EPO. It stores data about all patents, from 1978 to the current year, coming from about 90 patent offices worldwide, comprising the most relevant ones, such as EPO and USPTO.

As pointed out above, data is registered in PATSTAT in a raw format. To facilitate its analysis, ICRIOS processed it and produced a cleaned and harmonized database, i.e., PATSTAT-ICRIOS. This includes all bibliographic variables concerning each patent application. In particular, it stores application number and date, publication number and date, priority, title and abstract, application status, designed states for protection, main and secondary International Patent Classification (IPC) codes, name and address of both the applicant and the inventor, references (i.e., citations) to prior-art patent and non-patent literature, the corresponding Nomenclature of Units for Territorial Statistics (NUTS3) and, finally, File Index concordance tables, allowing the conversion of IPC codes into more aggregated and manageable technological classes.

To perform our investigation in the most correct and effective way, we carried out a pre-processing activity on the data of our interest. For this purpose, we used the framework R [7]. Our pre-processing activity consisted of the following tasks:

- *Data Extraction.* During this task, we first identified all the tables of PATSTAT-ICRIOS necessary for our analyses. To increase the effectiveness of the next tasks, we removed all the unnecessary and redundant attributes from these tables. This led to a strong reduction of the size of the data to process.
- *Data Normalization.* During this task, we removed some inhomogeneities regarding the data types of some fields (i.e., strings and dates).
- *Data Aggregation.* During this task, we performed a data integration activity aiming at storing all data about a concept in a unique collection.
- *Data Loading.* During this task, we loaded available data (represented in the CSV format) into a MongoDB [6] final database, which we used for our next activities.

At the end of these four tasks, the size of the dataset to analyze was reduced from 12.5 GB to 2.5 GB.

11.3.2 Support model

In this section, we introduce the data model representing data about patents and used by our approach. Before illustrating it, we must introduce two sets allowing us to formalize data at our disposal. These are: *(i)* the set Pat of all the patents stored in PATSTAT-ICRIOS, and *(ii)* the set Pat_k of the patents filed by at least one inventor of the country k.

We are now able to present our data model. It consists of a network $N = \langle V, A \rangle$. V denotes the set of the nodes (or vertices) of N. A node $v_i \in V$ corresponds exactly to a patent $p_i \in Pat$. Since there is a biunivocal correspondence between a node of V and the corresponding patent of Pat, in the following, in some cases, we adopt the symbol v_i to represent both of them and we adopt the terms "patent" and "node" interchangeably. Each node $v_i \in V$ has an associated label l_i, denoting the set of the countries of the inventors of p_i. A is the set of the arcs of N. There exists an arc $a_{ij} = (v_i, v_j) \in A$ if p_i cites p_j. Clearly, N is a directed network.

Starting from N, we can define some sets representing the neighborhoods of a node in V. In particular, given a node $v_i \in V$, we can define the following neighborhoods:

- $Cited_i$, i.e., the set of the patents cited by p_i:

  $$Cited_i = \{v_j | (v_i, v_j) \in A, v_j \neq v_i\}$$

  In other words, $Cited_i$ is the set of the nodes (and, therefore, the set of the patents) v_j such that there exists an arc from v_i to v_j (which implies that v_j was cited by v_i) in the set A of the arcs of N.
- $Citing_i$, i.e., the set of the patents citing p_i:

  $$Citing_i = \{v_j | (v_j, v_i) \in A, v_j \neq v_i\}$$

  In other words, $Citing_i$ is the set of the nodes (and, therefore, the set of the patents) v_j such that there is an arc from v_j to v_i (which implies that v_j cited v_i) in the set A of the arcs of N.
- V_k, i.e., the set of the nodes associated with the patents of Pat_k:

  $$V_k = \{v_i | v_i \in V, k \in l_i\}$$

  or, analogously:

  $$V_k = \{v_i | v_i \in V, p_i \in Pat_k\}$$

  In other words, V_k is the set of the nodes of N having the country k among the ones forming its label l. This is equivalent to say that V_k is the set of the patents having at least one inventor of the country k.

11.4 Centrality measures

11.4.1 Theoretical definition

The definition of the new centrality measure, well tailored for the patent scenario, represents the core of this paper. In fact, as pointed out in the Introduction, patent citations have a very important specificity because, if a patent p_i cites a patent p_j, the value of p_i decreases. As a consequence, differently from many other contexts, such as scientific papers, making a citation is not painless for the citing patent.

If we report this reasoning to our model, it implies that having incoming arcs is extremely positive for a node (and this is in line with the classic centrality metrics of network analysis). By contrast, having outgoing arcs is penalizing for a node (and this fact is not captured by classic centrality measures).

Since our support network is a directed one, it is necessary to define both the indegree and the outdegree of a node. The former indicates the number of its incoming arcs (i.e., the number of citations received by the corresponding patent), whereas the latter denotes the number of its outgoing arcs (i.e., the number of citations performed by the corresponding patent).

We propose two centrality measures, which we call:

- Naive Patent Degree (NPD);
- Refined Patent Degree (RPD).

We start by analyzing Naive Patent Degree. Given a node $v_i \in V$, the corresponding Naive Patent Degree NPD_i is defined as:

$$NPD_i = |Citing_i| - |Cited_i|$$

Clearly, this definition is immediate and captures the specificity mentioned above. However, we tried to find a more rigorous centrality metric, capable of capturing the synergies characterizing the patent scenario. Refined Patent Degree is the result of this effort. Its definition is based on the following considerations:

- C_1: given a patent p_i, the higher its capability of being cited by patents making very few citations, the higher its importance.
- C_2: given a patent p_i, the higher its capability of being cited by important patents, the higher, in turn, its importance. Observe that, in principle, Condition C_2 is very complex because it implies that the RPD of a node n_i depends on the RPD of a node n_j. This implies that, for the computation of this metric, complex systems characterized by hundreds, or even thousands, of equations and variables should be solved, at least in the most complex cases. As a consequence, the computation of RPD appears difficult to handle without a heuristic. A reasonable one could

consider the NPD of n_j, instead of the RPD of this node, in the computation of the RPD of n_i.

- C_3: the weight of a citation of a patent p_j, which a patent p_i must make, is inversely proportional to the number of citations received by p_j. In other words, if p_j is a very important patent, which received a very high number of citations, the fact that p_i must cite p_j does not considerably decrease the innovativity of p_i. By contrast, if p_i must cite a little cited patent p_j, it is possible to conclude that it is strongly influenced by p_j, and this significantly undermines its innovativity.

Taking all these conditions into account, RPD_i can be defined as:

$$RPD_i = \sum_{j=1}^{|Citing_i|} \omega_j - \sum_{q=1}^{|Cited_i|} \frac{1}{1+|Citing_q|}$$

where:

$$\omega_j = \alpha \left(\frac{1}{1+|Cited_j|}\right) + (1-\alpha)\left(\frac{NPD_j}{NPD_{max}}\right)$$

Here, $|Citing_i|$ (resp., $|Cited_i|$) is the cardinality of the set $Citing_i$ (resp., $Cited_i$). ω_j is a weighted mean of two terms. The former expresses Condition C_1, whereas the latter represents Condition C_2. The weight α allows the tuning of the mutual relevance of these two terms. In our case, we chose to assign the same importance to them; as a consequence, we set α equal to 0.5. Finally, the second term of the formula for RPD_i allows the formalization of Condition C_3.

As it will be clear in the next subsection, RPD does not overturn NPD. It simply refines this last metric, thanks to the three conditions, which it is based on. Specifically, it can produce acceptable distributions also for those countries having a low number of patents associated with them. This is exactly the scenario where NPD shows its main weaknesses.

11.4.2 Experimental evaluation

We started the evaluation of our metrics by computing the distribution of NPD for many world countries. Obtained results show that, for most countries, the distribution of NPD follows a power law. However, this power law is very singular and completely different from the ones generally characterizing degree distribution in network analysis.

In order to give an idea of the peculiarities of the distribution of NPD, in Figure 11.1, we show its values for Italy. From the analysis of this figure, we can see that, actually, there are two power law distributions almost mirrored with respect to the zero value of NPD.

Another interesting phenomenon, which can be observed in this figure, regards the two tails of the power law distributions. In fact, the right tail is much longer than the left one. This means that the number of citations received by Italian patents is much higher than the number of citations made by them. Furthermore, if we consider the shape of the tails, we can observe that the right tail is much steeper than the left one. This means that the distribution of citations received by Italian patents follows a more pronounced power law than the distribution of citations made by them. Finally, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.55.

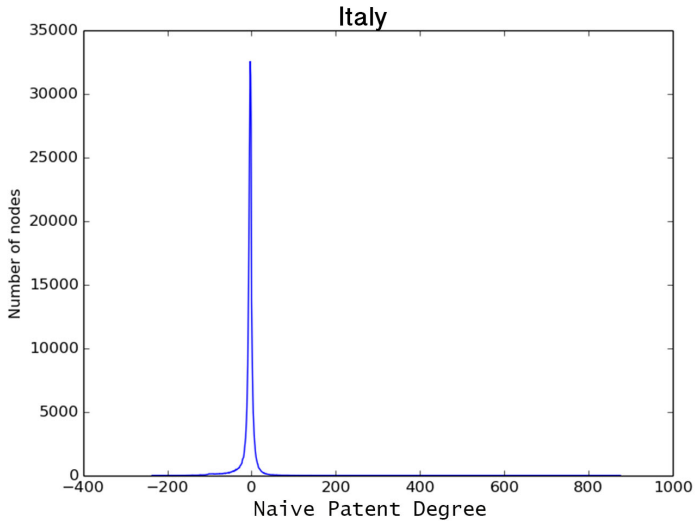


Fig. 11.1. Distribution of the values of NPD for Italy

As previously pointed out, the same trend (with the same specificities) can be observed for most countries.

For some countries, the distribution of NPD is similar to the one of Italy, even if much more disturbed than it. An example of this trend is shown in Figure 11.2, where we report the case of Estonia. A first result emerging from the comparison of this figure with Figure 11.1 is that the number of patents of Estonia is much lower than the one of Italy. Furthermore, we can note that, in this case, the trend of NPD values differs from the optimal one. This fact is more evident in the left power law distribution. Here, it is possible to observe some peaks that evidence the presence of a considerable number of Estonian patents that make many citations, especially if we compare their number with the total number of Estonian patents. As a further result, we observe that the length of the right and the left tails are comparable. However, also in this case, the right tail is steeper than the left one. All the previous observations are valid for all the countries with such a kind of trend for NPD. In this case, the

ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 1.05.

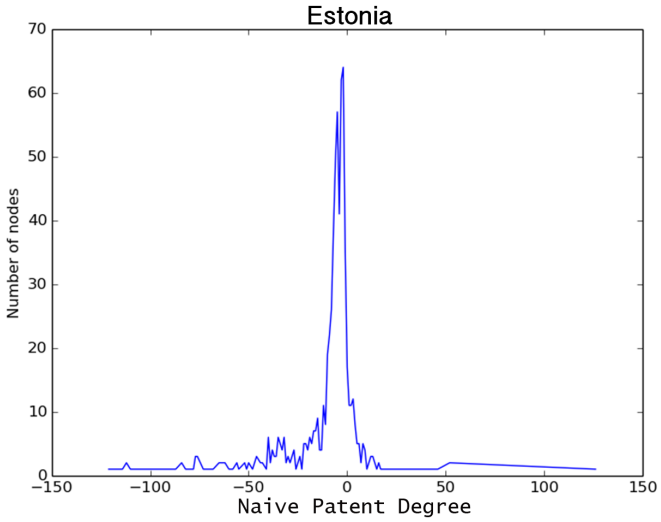


Fig. 11.2. Distribution of the values of NPD for Estonia

For some countries, the distribution of NPD does not follow a power law. As an example of this situation consider Figure 11.3, where we report the distribution of NPD for Tunisia. In this figure, we can also observe that the left tail is longer than the right one and that the number of Tunisian patents is very low. Even if this case is not very significant from a statistic point of view, we can again observe that the right "tail" is "steeper" than the left one. Furthermore, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 2.64. This also happens for the other countries with an analogous distribution of NPD.

The comparison of the results obtained for the three kinds of country mentioned above suggests that the most innovative and rich countries present a power law distribution for NPD. Furthermore, since these countries drive the innovation and the technological progress of the other ones, their patents receive many more citations than the ones they must make.

Those countries, like Estonia, showing a disturbed power law for NPD do not have a patent patrimony allowing them to be innovation leaders currently. However, they are accumulating a certain number of patents allowing them to become innovation leaders in the near future.

Finally, those countries, like Tunisia, having an irregular distribution of NPD are characterized by a very low number of patents. They have not reached an adequate

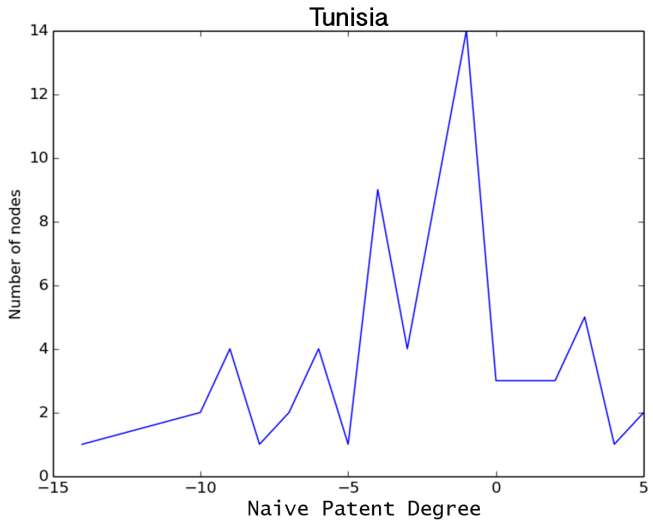


Fig. 11.3. Distribution of the values of NPD for Tunisia

research and innovation level yet. Their very limited number of patents does not allow a detailed analysis about their situation.

After having evaluated NPD, we proceed to investigate RPD. We start with the most innovative countries. In Figure 11.4, we report the distribution of the values of RPD for Italy on the left, and a zoomed representation of the same distribution around the zero value of RPD on the right. If we compare the distribution of RPD with the one of NPD, reported in Figure 11.1, we can observe that RPD confirms (or, even better, magnifies) all the results returned by NPD. The only exception regards the steepness of the two tails. In fact, differently from NPD, in this case, the left tail is steeper than the right one. Finally, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.14.

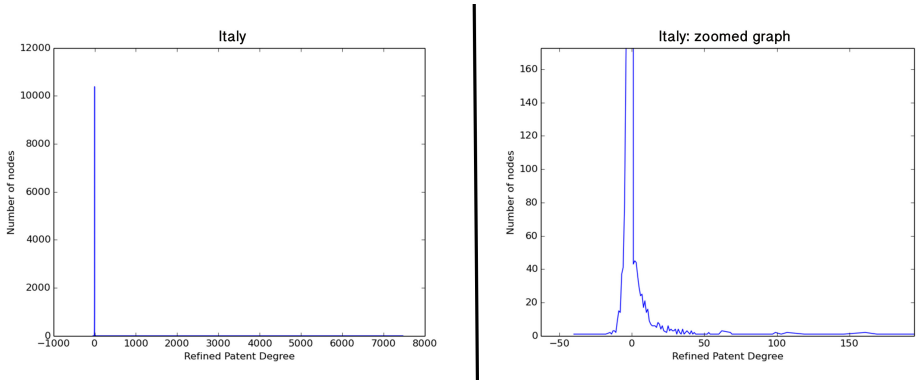


Fig. 11.4. Distribution of the values of RPD for Italy

In Figure 11.5, we report the distribution of the values of RPD for Estonia, as a representative of the countries with an intermediate number of patents. If we compare this distribution with the corresponding one of NPD for the same country, we can observe that RPD removes many of the disturbances observed in NPD. Therefore, the corresponding distribution is much "cleaner". Differently from what happens in Figure 11.4, and analogously to the trend shown in Figure 11.2, we have that, in this case, the right tail is steeper than the left one. In this case, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.20.

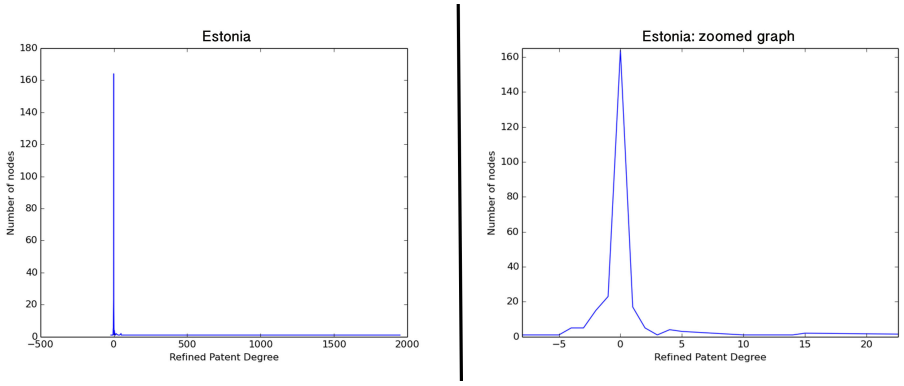


Fig. 11.5. Distribution of the values of RPD for Estonia

An analogous reasoning can be drawn for those countries having a low number of patents. If we compare the distribution of RPD for Tunisia, shown in Figure 11.6, with the corresponding one of NPD, shown in Figure 11.3, we can see that the RPD's capability of cleaning the distortions of NPD is even magnified for countries with a small number of patents. In this case, the steepness of the left tail is slightly higher than the one of the right tail, even if the differences are not remarkable. Furthermore, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.33.

In conclusion, both NPD and RPD appear well suited as centrality measures for patents. However, RPD is capable of removing some distortions that have been shown by NPD when this last is adopted for evaluating countries with a small number of patents.

To make our analysis about NPD and RPD more exhaustive, we computed the "similarity rate" of the results returned by NPD and RPD. For this purpose, given a country k, we computed the set Top_k^{NPD} (resp., Top_k^{RPD}) of the top 5% of the patents of Pat_k with the highest values of NPD (resp., RPD). Then, we computed the parameter:

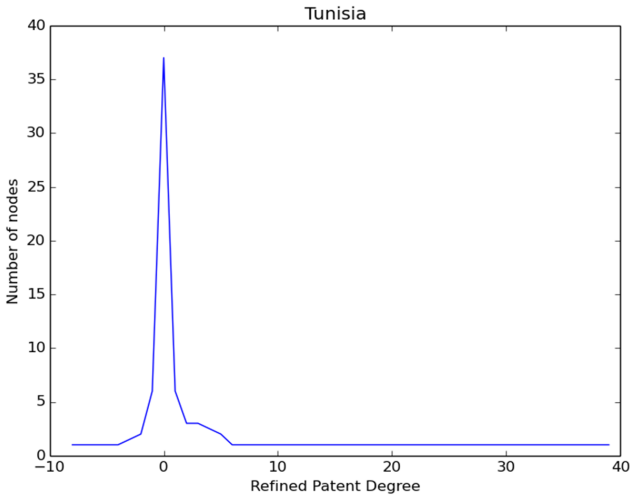


Fig. 11.6. Distribution of the values of RPD for Tunisia

$$rTop_k = \frac{|Top_k^{NPD} \cap Top_k^{RPD}|}{|Top_k^{NPD}|}$$

The possible values of $rTop_k$ range between 0 and 1, where 0 denotes that NPD and RPD return completely different results, whereas 1 indicates that they have exactly the same behavior.

We computed the value of $rTop_k$ for the world countries and, in Table 11.1, we report some of them. From the analysis of this table, we can observe that the value of $rTop_k$ is generally much higher than 0.5. Its average value for all world countries is 0.65. This result, along with the previous ones specified above, allows us to conclude that RPD does not overturn NPD. Actually, the former refines the latter thanks to the three conditions, which it is based on. RPD can return acceptable and clean distributions also for those countries having a low number of patents, in which case NPD is excessively sentitive to disturbances.

11.5 Some possible applications

Our new patent centrality measures can have a lot of applications. In order to give an idea of them, in this section, we present three applications, namely: *(i)* the computation of the "scope" of a patent; *(ii)* the definition of the lifecycle of a patent; *(iii)* the detection of "power patents".

11.5.1 Computation of the scope of a patent

We use the term "scope" to indicate the width and the strength of the influence of a patent $p_i \in Pat$ on the other patents, that is the width and the strength of the influence of a node $v_i \in V$ on the other nodes of N. We argue that the scope of v_i is strictly

Country	$rTop_k$
Algeria	1.00
Austria	0.86
Brazil	0.62
Bulgaria	0.68
China	0.56
South Corea	0.62
Denmark	0.59
Estonia	0.77
Finland	0.52
France	0.57
Germany	0.65
Japan	0.73
Greece	0.50
India	0.61
Italy	0.59
Luxembourg	1.00
Polan	0.63
United Kingdom	0.59
Romania	0.67
Russia	0.59
Spain	0.48
South Africa	0.57
Taiwan	0.60
Tunisia	0.67

Table 11.1. Similarity Rate of NPD and RPD for some countries

connected to the number and the centrality of the nodes citing it, either directly or indirectly. As a consequence, in the scope definition, the main roles are played by the centrality measure, which we have already seen, and by the neighborhood of a node, which we introduce now.

With regard to this last concept, we point out that there could exist several levels of neighborhood of a node v_i. For this reason, it is possible to introduce the neighborhood of level t of a node $v_i \in V$. This is defined as follows:

$$nbh_i^t = \begin{cases} Citing_i & if\, t = 0 \\ \{v_j | (v_j, v_l) \in A, v_l \in nbh_i^{t-1}\} & if\, t > 0 \end{cases}$$

We are now able to define the Naive Scope NS_i^t and the Refined Scope RS_i^t of a node $v_i \in V$ w.r.t. the nodes of its t^{th} neighborhood nbh_i^t as follows:

$$NS_i^t = \sum_{j \in nbh_i^t} NPD_j \qquad RS_i^t = \sum_{j \in nbh_i^t} RPD_j$$

Once the scope of a node has been defined, it is possible to perform an investigation at the country level to analyze the average trend of the scope of the nodes of a country k. In particular, the Average Naive Scope ANS_k^t and the Average Refined Scope ARS_k^t of the patents of a country k with respect to their t^{th}-level neighbors can be defined as:

$$ANS_k^t = \frac{\sum_{v_i \in V_k} NS_i^t}{|V_k|} \qquad ARS_k^t = \frac{\sum_{v_i \in V_k} RS_i^t}{|V_k|}$$

We computed the trends of ANS_k^t and ARS_k^t for most world countries. As an example, in Figures 11.7 - 11.9, we show the trend of ANS_k^t (in blue) and ARS_k^t (in red) for three countries, namely China, Luxembourg and Poland. Analogous trends have been found for the other countries. From the analysis of Figures 11.7 - 11.9, we can observe that, for all cases, the average scope decreases when the neighborhood level increases. This general result was expected. However, the really interesting analysis concerns *how fast* this decrease is. As for this issue, we generally observe a steep decrease so that, after the third-level neighborhoods, patent scopes are almost null. If we compare the trends of ANS_k^t and ARS_k^t in these figures, we can observe that they are similar, even if the trends of ARS_k^t are always steeper than the ones of ANS_k^t. This is in line with the results of the comparison of NPD and RPD presented in Section 11.4.2, where we have seen that RPD refines and magnifies the trends characterizing NPD.

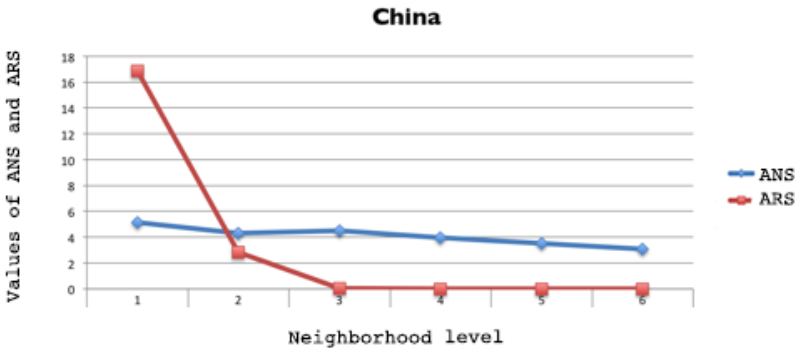


Fig. 11.7. Trend of ANS_k^t and ARS_k^t against the neighborhood level t for China

11.5.2 Computation of the lifecycle of a patent

This activity aims at verifying if, by computing, year by year, the NPD and the RPD of patents published all over the world, it is possible to determine one or more characteristic patterns. In the affirmative case, each characteristic pattern would represent a lifecycle template for the patents following it. Defining lifecycle templates for specific categories of patents is extremely useful because, given a new patent p_i belonging to a category for which there exists a lifecycle template, it is possible to foresee the NPD and the RPD of p_i over time, and, ultimately, the number and the relevance of the citations received by it.

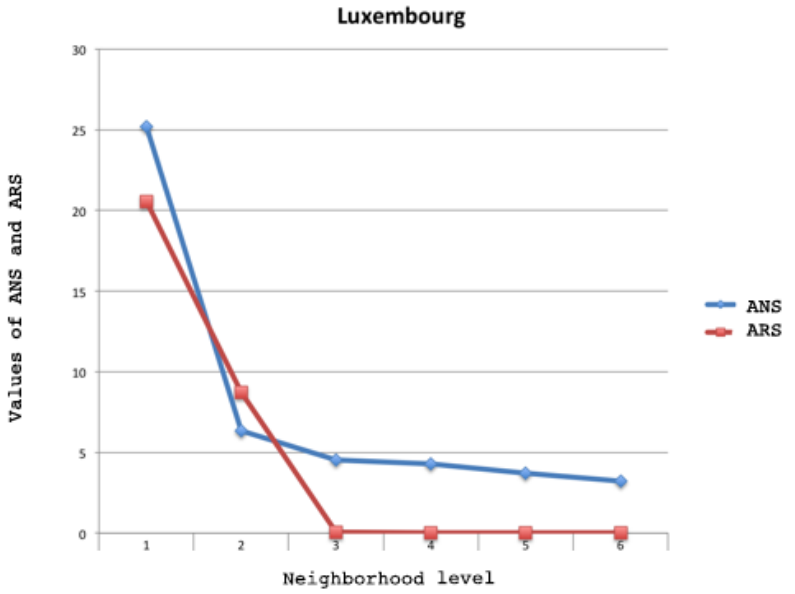


Fig. 11.8. Trend of ANS_k^t and ARS_k^t against the neighborhood level t for Luxenbourg

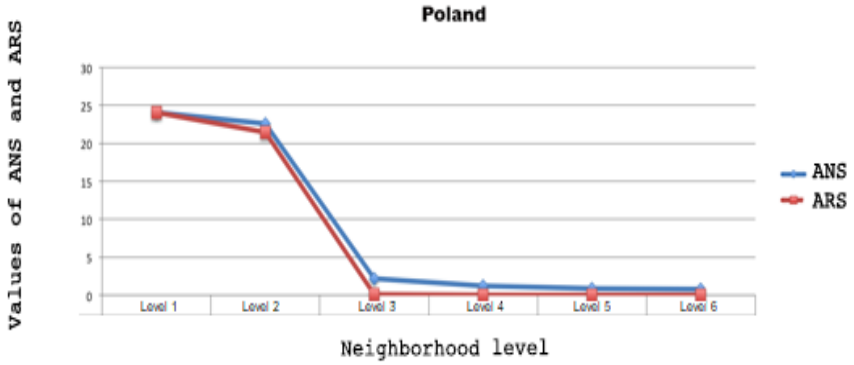


Fig. 11.9. Trend of ANS_k^t and ARS_k^t against the neighborhood level t for Poland

In order to show how lifecycle templates could be defined, in the following, we associate categories with years and introduce a category per year. However, we could adopt the same technique with a completely different taxonomy, for instance by associating a category per IPC class (in such a way as to define a patent lifecycle template for each IPC class), a category per country, and so forth.

To construct a lifecycle template for each year, we must preliminarily introduce the measures NPD_i^y and RPD_i^y. These two measures are analogous to NPD_i and RPD_i, except that they consider only the patents published in the year y.

To carry out our analysis, for each year from 1985 to 2013, we considered all the patents published in that year and, for each of them, we computed the values of NPD and RPD from that year until 2013. For instance, in Figure 11.10 (resp., 11.11,

11.12 and 11.13), we show the trends of RPD for the patents published in the year 1985 (resp., 1990, 1995 and 2000). By analyzing the obtained results we have seen that, independently of the publication year of patents, there exists a unique pattern representing the patent lifecycle.

We aimed at expressing this lifecycle template matematically and we observed that it can be represented by a sixth-degree polynomial function of the form:

$$y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + e$$

To give a visual intuition of this fact, in Figures 11.10 – 11.13, we traced, along with the real values of patent lifecycle, the sixth-degree polynomial function that best approximates it. It is possible to observe that the deviations between the real values and the ones of the polynomial function are very small.

By analyzing each figure, we can observe that RPD is negative in the publication year of patents. This is due to the fact that all the citations performed by a given patent p_i are concentrated in its publication year, whereas, in that year, no patents, or a little number of them, cite p_i. After the first year from the publication of p_i, the corresponding RPD starts to increase. This increase reaches a maximum after about 5 years from publication. Then, a stall phase can be observed until to about the eighth year; this phase is followed by a phase of decline, which becomes stronger and stronger until the RPD of p_i reaches an almost null value. This decline can be easily explained by considering that, for most patents, after about ten years from their publication, new technologies and/or more innovative patents appear, which make them obsolete.

In Table 11.2, we report the values of the coefficients of the sixth-degree polynomial function that represents the lifecycle templates regarding patents published in the years 1985-2000, obtained by applying the least square method. The coefficients of the lifecycles regarding patents published after 2000 are not reported because these lifecycles are too recent and, consequently, they are not complete yet.

Very similar trends and conclusions can be derived for NPD.

11.5.3 Definition of power patents and investigation of their importance

The definition of patent-tailored centrality measures like ours allows the identification of the most relevant patents. As a matter of fact, since both NPD and RPD follow a power law, it is reasonable to assume that there exist some *power patents*, i.e., a very small number of patents that have been cited very much. In order to investigate this aspect, in the following, we will consider RPD, even if analogous reasonings can be made for NPD.

Clearly, in principle, the fraction of power patents could differ for each country because it depends on the trend of the corresponding distribution of the RPD values.

Years	*a*	*b*	*c*	*d*	*e*	*f*	*g*
1985	-1E-06	1E-04	-0,0039	0,0778	-0,8166	3,9637	-3,9546
1986	-1E-06	0,0001	-0,0046	0,0890	-0,8942	4,1551	-3,9498
1987	-2E-06	0,0002	-0,0056	0,1033	-0,9902	4,4030	-4,0791
1988	-2E-06	0,0002	-0,0066	0,1171	-1,0779	4,6154	-4,0942
1989	-3E-06	0,0002	-0,0078	0,1312	-1,1494	4,7282	-4,1012
1990	-3E-06	0,0003	-0,0084	0,1350	-1,1406	4,5921	-3,9704
1991	-4E-06	0,0004	-0,0113	0,1668	-1,3066	4,9941	-4,4076
1992	-5E-06	0,0005	-0,0118	0,1768	-1,4087	5,2030	-4,7034
1993	-8E-06	0,0006	-0,0154	0,2149	-1,5778	5,6661	-4,9479
1994	-1E-05	0,0008	-0,0198	0,2619	-1,8236	6,2006	-5,1879
1995	-1E-05	0,0009	-0,0225	0,2841	-1,8956	6,2383	-5,2146
1996	-2E-05	0,0011	-0,0260	0,3124	-1,9979	6,3822	-5,4142
1997	-2E-05	0,0014	-0,0305	0,3474	-2,1273	6,5878	-5,6143
1998	-3E-05	0,0014	-0,0306	0,3380	-2,0453	6,3775	-5,5960
1999	-3E-05	0,0016	-0,0341	0,3659	-2,1663	6,6400	-5,8417
2000	-4E-05	0,0020	-0,0393	0,4066	-2,3270	6,9163	-6,1626

Table 11.2. Values of the coefficients of the sixth-degree polynomial function that best approximates the lifecycles of patents published from 1985 to 2000

However, thanks to the features of RPD illustrated in Section 11.4.1, if we choose to select as power patents those ones whose values of RPD lie at the right of the elbow of the RPD distribution function, we obtain that, for most countries, it is sufficient to take as power patents the top 5% of patents having the highest values of RPD. To give an idea of this reasoning, in Figures 11.14, 11.15 and 11.16, we show three examples concerning the RPD value distribution of India, France and Japan. In all the three cases, it is evident that taking as power patents the top 5% of patents is sufficient. Analogous trends can be found for almost all the other world countries. In the following, we indicate with $\overline{Pat_k}$ the power patents of the country k.

After having defined a way to detect the power patents of each country, we aimed at investigating if, for a patent p_j, being cited by a power patent p_i can bring benefits, i.e., citations performed by patents that, having cited p_i, must also cite p_j.

To answer this question, we must preliminarily introduce some parameters. In particular, let $p_i \in Pat_k$ be a patent of the country k:

- The set of potential beneficiaries PB_i of p_i is defined as:

$$PB_i = \{p_j | p_j \in Cited_i, p_i \in Cited_r, p_j \in Cited_r\}$$

- The fraction of potential beneficiaries of p_i is defined as:

$$F_i^{PB} = \frac{|PB_i|}{|Cited_i|}$$

- The average fraction of the potential beneficiaries of the patents of a country k is defined as:

$$AvgF_k^{PB} = \frac{\sum_{p_i \in Pat_k} F_i^{PB}}{|Pat_k|}$$

- The average fraction of the potential beneficiaries of the power patents of a country k is defined as:

$$\overline{AvgF_k^{PB}} = \frac{\sum_{p_i \in \overline{Pat_k}} F_i^{PB}}{|\overline{Pat_k}|}$$

We are now able to define the benefit capability bc_k of the power patents of a country k. Specifically:

$$bc_k = \frac{\overline{AvgF_k^{PB}}}{\overline{AvgF_k^{PB}}}$$

The value of bc_k ranges between 0 and $+\infty$. If $bc_k \leq 1$, the power patents of k do not provide benefits to the patents cited by them. By contrast, if $bc_k > 1$, they are beneficial for the patents cited by them, and the higher bc_k the greater these benefits.

In Table 11.3, we report the value of bc for several countries. From the analysis of this table, we can see that bc is generally much higher than 1. This clearly evidences that, for a patent, obtaining a citation from a power patent is highly beneficial.

Country	*bc*
Austria	10.73
Brazil	0.47
China	13.30
South Korea	17.23
Denmark	6.58
Finland	7.93
France	10.37
Germany	9.72
Japan	5.19
Greece	1.47
India	21.63
Italy	10.11
Poland	6.32
United Kingdom	4.98
Romania	12.46
Russia	21.23
Spain	12.36
South Africa	6.24
Taiwan	17.73

Table 11.3. Values of bc for several countries

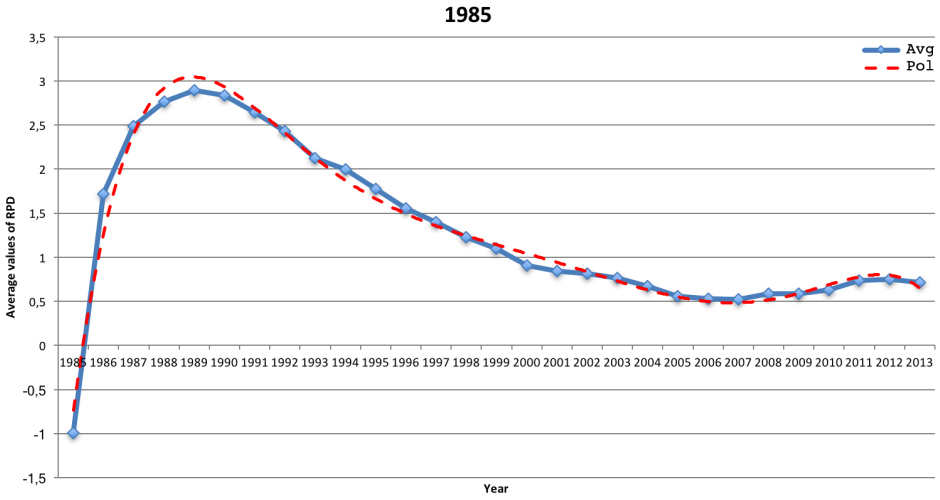


Fig. 11.10. Average values of RPD over time for the patents published in 1985

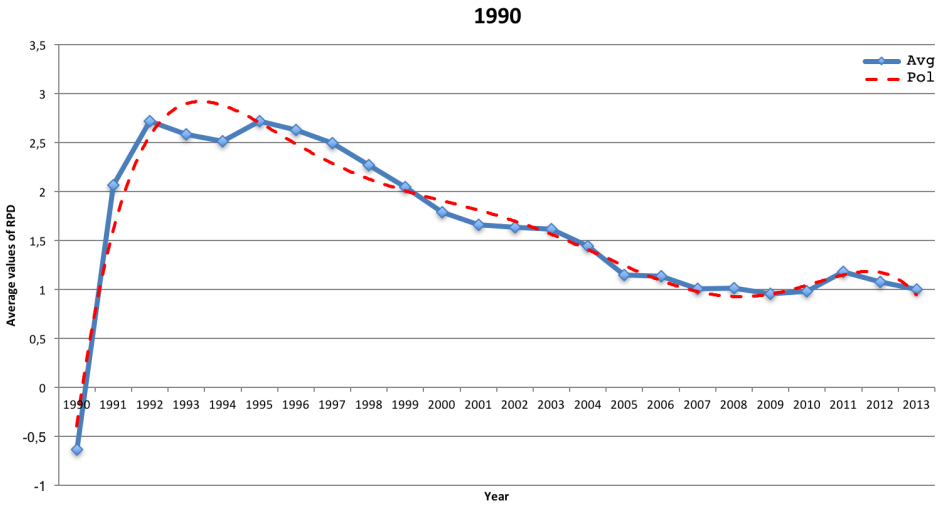


Fig. 11.11. Average values of RPD over time for the patents published in 1990

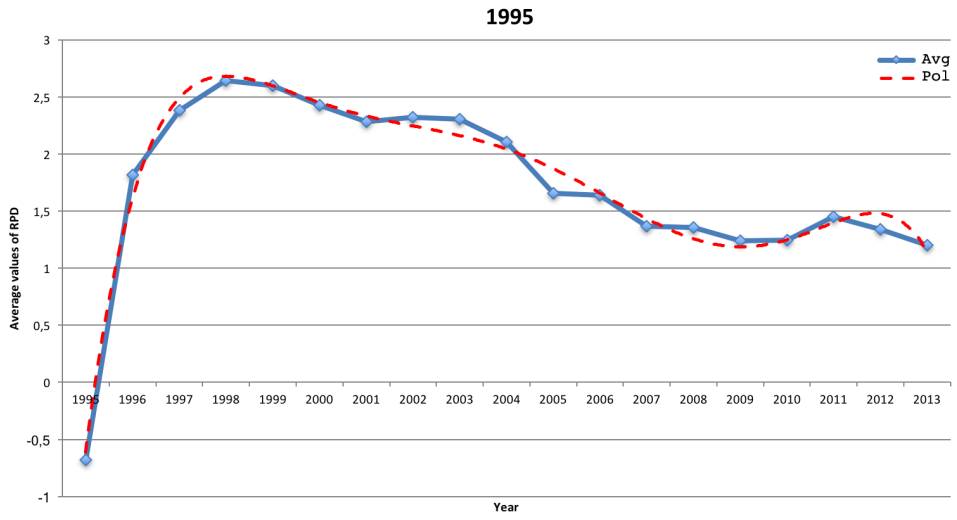


Fig. 11.12. Average values of RPD over time for the patents published in 1995

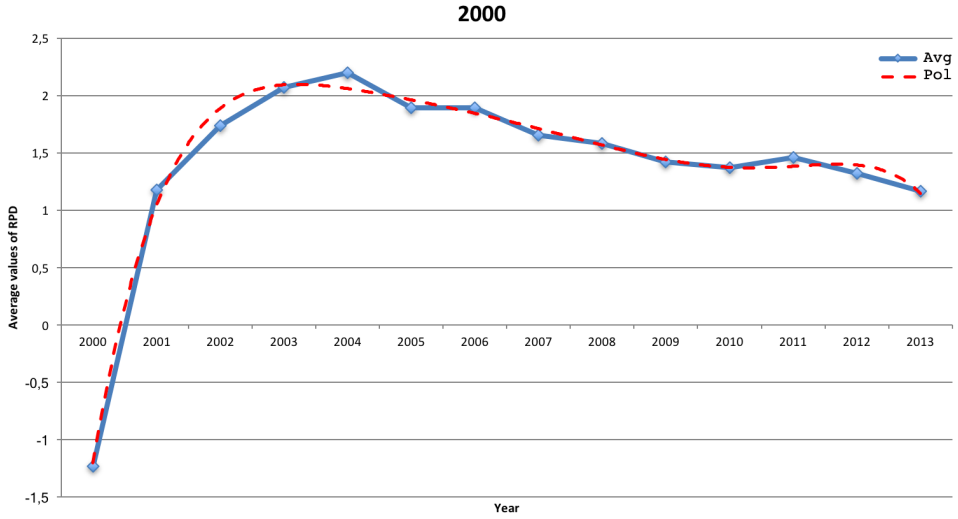


Fig. 11.13. Average values of RPD over time for the patents published in 2000

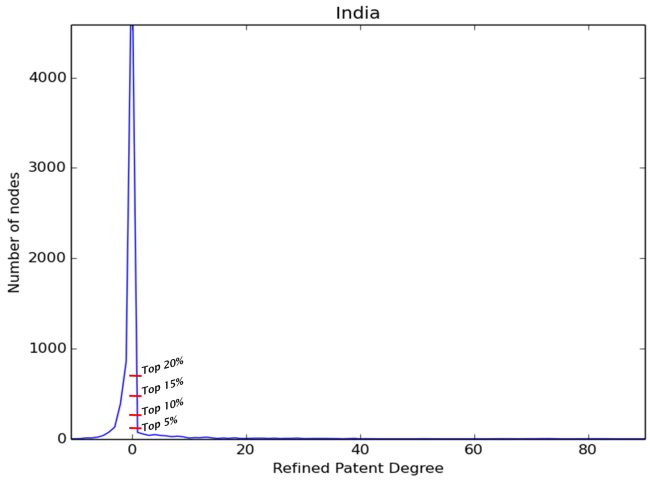


Fig. 11.14. Distribution of the values of RPD for India, along with the levels corresponding to the top 5%, 10%, 15% and 20% of patents with the highest values

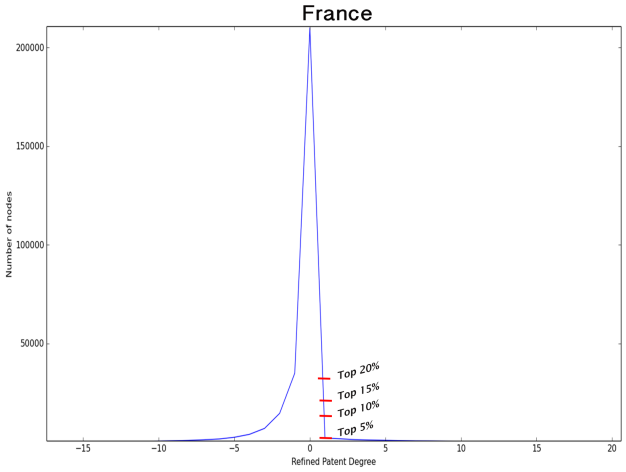


Fig. 11.15. Distribution of the values of RPD for France, along with the levels corresponding to the top 5%, 10%, 15% and 20% of patents with the highest values

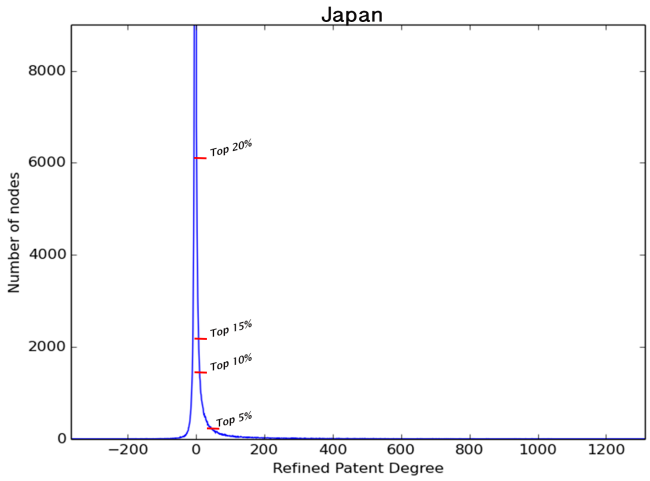


Fig. 11.16. Distribution of the values of RPD for Japan, along with the levels corresponding to the top 5%, 10%, 15% and 20% of patents with the highest values

12

Extraction of Knowledge Patterns

12.1 Introduction

In the last years, scientometrics and bibliometrics received a growing interest both in research literature and as objective ways for evaluating the performances of researchers, universities, institutions, etc. Indeed, research collaborations across institutions, firms and countries have been largely investigated in strategy and management literature [409, 308, 79, 318, 170, 344, 173, 383, 379, 121]. Moreover, different studies have been performed to understand whether international flows from developed countries to developing and less-developed ones have some positive effects in these last ones [178]. Furthermore, many studies investigate the impact and the effects of international knowledge flows by focusing on R&D collaborations and inventions and on their impact on innovation [260, 300, 163, 177, 410, 78, 29, 16, 251, 236, 273, 285].

Currently, data available for scientometrics and bibliometrics investigations are growing at a very rapid rate. As a matter of fact, the problem of extracting useful knowledge from these data can be seen as a Data Mining problem, and in the very next future, big data approaches for solving it will be unavoidable. The obvious consequence of this reasoning is that more and more innovative approaches to addressing this issue are necessary.

Social Network Analysis [458, 53, 52, 24, 106, 107, 258, 328] and, more in general, graph theory, have been a prominent family of approaches adopted in the past in this context (see, for istance [276, 36, 46, 72, 446, 359, 10, 13, 277, 237, 103, 71, 11]). Furthermore, it is possible to foresee that they will be even more employed in the future, due to the more and more increasing number of proposals someway involving them.

All these studies have certainly contributed to a development of the research in innovation dynamics. However, there are still several aspects that need to be deepened. For instance:

- Most of these approaches focus on authors, whereas investigations on institutions would be extremely interesting. This fact is also valid for the paper that, to the best of our knowledge, is the only one analyzing hubs in the past [36]. In fact, in this chapter, the definition of hub is centered on authors.
- Most of the previous approaches employed only centrality measures in their analysis, whereas, in Social Network Analysis, there are several other parameters (e.g., the connection level of a network), which are at least as important as centrality.
- The past approaches did not investigate the neighbors of authors or institutions, whereas we know that, owing to the concept of homophily (that is a key concept in Social Network Analysis), the neighbor of a node can strongly influence the behavior of the corresponding author or institution.

This chapter aims at providing a contribution in this setting. Indeed, it proposes a new Social Network Analysis-based (hereafter, SNA-based) approach to extracting knowledge patterns about research activities and hubs in a set of countries of interest. As for this chapter, a hub is a research institution that operates as a guide or stimulus to the research in its country and, at the same time, is capable of stimulating cooperations with institutions of other countries. Our hub definition is strongly fitted to our scenario of interest. It strongly benefits from the observations, suggestions and experience of innovation management researchers, who guided us in its formulation.

Our approach is general and can be directly applied to any set of countries. The only requirement is to have at disposal the set of the publications of all the research institutions of the countries to investigate. In this chapter, we applied it to four North African countries (e.g., Algeria, Egypt, Morocco and Tunisia) and we used all the publications of all the research institutions of the four countries of interest in the time interval $[2003, 2013]$, as stored in the Web of Science repository [4].

The most important support data structure (already introduced in the past literature) is a social network with nodes that represent institutions and with edges that denote collaborations among institutions. Starting from it, other important support data structures and accompanying parameters (some of which were never defined in the literature) are introduced.

Thanks to our approach, it is possible to reconstruct a very detailed and multidimensional picture of the research scenarios of a set of countries, as well as to determine analogies and differences among them. In this way, innovation managers have at their disposal some empirical instruments helping their decisions. Beside providing several knowledge patterns about institutions and their collaboration, not known in the past, this chapter provides several other contributions and, in our opinion, some of them are even more important than the extracted knowledge patterns. Indeed:

- it presents a general SNA-based approach that can be applied to extract knowledge about research scenarios and the corresponding institutions for a set of countries;
- it redefines some SNA metrics in such a way as to make them suitable for this application scenario;
- it defines new metrics about institutions and their cooperations not presented in the past;
- it introduces the concept of hub and provides a method to determine the hubs of each country, as well as to investigate their main features;
- it defines new data structures (such as the *clique social network* and the *nbh social network*) allowing the extraction of interesting knowledge about hubs and their neighbors;
- it provides both a visual and a quantitative method to determine the core hubs (if they exist) of a given country.

For an expert, the extracted knowledge patterns are important for at least two reasons. First of all, they may improve her understanding of the impact of different socio-economic conditions on the structure and evolution of scientific collaborations. In this sense, the four countries, which our approach was applied on, present a great heterogeneity along several socio-economic dimensions, such as type and degree of economic specializations, language and culture. Secondly, this analysis may help the design of policy interventions aimed to sustain the accumulation of scientific and technological capabilities in the countries on which our approach is applied. For instance, the identification and analysis of hubs and their interactions with local research communities may lead to the design of policies that explicitly target hubs as key vectors to access and disseminate knowledge from advanced countries.

The algorithms implementing our approach are in Python [2] and the underlying DBMS is MongoDB [6]. As a consequence, our approach is already compliant with big data technology and, therefore, can help very large investigations (for instance, a large set of countries, or countries having a very high number of research institutions and publications, like United States and European countries).

this chapter is organized as follows. In Section 12.2, we present related literature. In Section 12.3, we describe available data and illustrate the pre-processing activities performed on them. In Section 12.4, we present our approach. In Section 12.5, we apply it to the four North African countries mentioned above. In Section 12.6, we illustrate the main novelties of our approach w.r.t. the related ones and we compare it with three commercial systems, i.e., Elsevier Pure, Scopus and Fingerprint Engine. Finally, in Section 12.6, we draw our conclusions and overview some possible future developments.

12.2 Related Literature

Research collaborations across firms and countries have been largely investigated in strategy and management literature. In this field, authors showed that these collaborations play a key role in the acquisition of external knowledge [409, 474] and in the creation of new knowledge [308, 382].

Specifically, in [308], the authors show that cross-regional networking positively influences innovation, at least in Europe. However, they also show that regional labor mobility plays an even more relevant role. In line with the approach adopted in [308], [79] investigates patent application in biotechnology, organic chemistry and pharmaceutical, and shows that network activity across firms and location is extremely important in the localization of knowledge flows. In [474], the authors employ data mining techniques to show that local research groups, characterized by a very high internal cohesion, hinder knowledge transmission. At the same time, they show that scientists with a centralized position in a network have a positive effect on knowledge flow. [318] analyzes the interactions among researchers coming from developing and advanced countries and finds that innovation in Latin American countries was largely influenced by R&D activities carried out on some OECD countries. [170] investigates cross-border inventions between BRICS firms and European Union actors and finds that these inventions are growing more valuable than the domestic ones. [344] studies and analyzes some survey interviews about Nigerian firms and employs them to determine which factors guide these firms to successfully or unsuccessfully adopt industrial innovations. [173] investigates the learning processes and the linkage behavior of small and large, local and foreign firms in Tanzania. [383] analyzes the efficiency of South Africa's innovation system. [379] investigates innovation in Ghana through a multi-level theoretical framework. In [121], the author proposes a framework aimed to evaluate the optimal conditions for innovation in emerging economies, with a special focus on Kenya and Uganda. The paper shows that, in both countries, the human capital and the firm's internal infrastructure play a significant role in innovation.

Another important investigation in innovation is aimed to understand whether international flows from developed countries to developing and less-developed ones have some positive effects in these last countries. The role played by knowledge spillovers is well known in the literature (see, for instance, [178]). These spillovers can operate through many channels, ranging from formal communication methods (e.g., scientific publications) to informal ones (e.g., person-to-person contacts).

In [382], the authors analyze the institutions having the highest impact on collaboration networks. Their researches confirm the existence of elite groups that cooperate

with other minor institutions. This form of cooperation allows the creation and diffusion of knowledge on management.

Many studies investigate the impact and the effects of international knowledge flows by focusing on international R&D collaborations and inventions and on their impact on innovation. This investigation was performed at two different levels (i.e., theoretical and empirical ones). From a theoretical point of view, some authors argued that these collaborations can lead to higher-quality innovations, thanks to the contamination of different skills and pieces of knowledge [260, 300]. Other authors hypothesized that international collaborations are not efficient owing to high coordination costs and difficulties to integrate knowledge of different research teams [163, 177, 410]. Empirical studies produced mixed results. In fact, [78] found that, as far as Indian and Chinese inventors are concerned, cross-border inventions receive more citations than the ones produced by the inventors of only one country. In [29], the authors show that international collaborations positively influence patent quality; at the same time, they evidence the difficulties of research teams to absorb external knowledge. In [16], the authors show that research collaboration in Africa presents an inhomogeneous structure. They also evidence that these collaborations are strongly constrained by several factors, ranging from geography to history, culture and language. In [251], the authors conduct a study on North African countries. They evidence that, in these countries, research collaborations are rapidly changing although they are still weak. In [236], the authors analyze the international transmission of knowledge in USA. In [273], the authors present a deep research about the geography of innovation, based on patent analysis. They show how localized knowledge flows are largely mediated by labor and technology markets. In [285], the author shows that both the links and the h-indexes of co-inventors and co-authors highly enhanced the flows of academic knowledge into industrial patents in South Africa's firms, as well as knowledge diffusion in large R&D and innovation clusters and hubs. In [43], the authors investigate the diffusion of European knowledge. Specifically, they analyze the diffusion of knowledge between European countries and European Neighboring Countries (ENCs). For this purpose, they use several indicators allowing them to evaluate how European knowledge is employed by ENCs. Obtained results show that ENCs can benefit from the interaction with European countries and can "transform" European knowledge and tools in new knowledge and innovation.

Social Network Analysis and, more in general, graph theory have been largely employed to investigate co-authorship networks and research scenarios in the past. The structure of co-authorship networks in three different fields (i.e., Nanoscience, Pharmacology and Statistics) in Spain in the time interval $[2006, 2008]$ is analyzed in [71]. Here, the authors investigate if there exists a relationship between the research

performance of authors and their position in co-authorship networks. A co-authorship network in the interdisciplinary field of "evolution of cooperation" is analyzed in [276]. To carry out their investigation, the authors adopt SNA and a modularity measure. A co-authorship network regarding Digital Libraries is investigated in [277]. For this purpose, some support social networks are constructed and analyzed to determine the impact of authors in the co-authorship network. To evaluate this impact, several SNA measures, such as centrality and PageRank, along with a new metric, called AuthorRank, are employed. In [36], the authors investigate co-authorship networks involving four institutions to understand how information flows therein and to detect the leader authors (called hubs). Hubs are defined as those authors having both a high eigenvector centrality and a high betweenness centrality. Hub detection is performed by means of SNA-based techniques. After the detection of the hubs of the four considered institutions, the authors analyze the relationships among them. In [46], a co-authorship network is analyzed to understand if it is possible to link centrality measures with author performances and if the authors' gender can have an impact on their performance. A co-authorship network, constructed starting from the publications in "information visualization" field in the time interval $[1974, 2004]$, is investigated in [72]. In [446], the authors hypothesize that international cooperation networks are self-organizing. To verify their hypothesis, they employ SNA-based techniques, capable of analyzing the growth of these networks. A method for building link predictors in networks with nodes that represent researchers and with links that denote collaborations is proposed in [359]. SNA-based techniques are employed in [10] for examining the effect of social networks on the performance of scholars in a given discipline. A co-authorship network concerning the "industrial ecology" field is investigated in [237], with the goal of evaluating the corresponding research efforts and results. Blockmodeling techniques are employed in [103] for analyzing a co-authorship network. In [11], the authors investigate whether preferential attachment in scientific co-autorship networks is different for authors with different forms of centrality. An exploratory analysis of co-authorship in the field of management and organizational studies is presented in [13]. Here, the authors determine the frequency of collaboration in the most prominent journals in the field. In [233], the authors apply classical SNA-based techniques on a complex co-authorship network to find knowledge patterns about paper citation, cooperation trends, the evolution of key components and author ranking. Furthermore, they employ the network diameter, clustering coefficient and degree distribution to find connectivity patterns, small-world network phenomena, and several other properties. In [125], the authors analyze a set of shared papers by constructing a two-mode network with node that represent both authors and papers, and by applying new regression models on it. After this, they employ the obtained

knowledge to perform an empirical analysis on a larger co-authorship network. In [139], the authors apply SNA-based techniques on a co-authorship network for estimating cooperation trends and for identifying the most important scientists and institutions. Furthermore, they investigate the possible application of their approach and of the derived knowledge to a medical context. In [99], the authors start from a demographic analysis to provide an overview of the corresponding distribution of both scientific labels and academic titles. In particular, they employ Social Network Analysis to investigate a co-authorship network and several citation metrics.

12.3 Available data and preprocessing

The dataset we used was derived from Web of Science of Thomson Reuters. It stores all the publications performed by all the research institutes of the four countries into examination from 2003 to 2013. It consisted of four parts concerning:

- *Institutions*. This part contains information about all the research institutions of the four countries into consideration, as well as about the research institutions of the other countries cooperating with them from 2003 to 2013.
- *Authorships*. This part contains information about all the authorships concerning papers involving at least one of the institutions into consideration from 2003 to 2013.
- *Publications*. This part stores information about the publications of the authors affiliated to at least one of the institutions into consideration.
- *Research areas and fields*. This part stores information about the research areas and fields, as classified by Web of Science.

A first analysis of our dataset allowed us to verify that the parts concerning institutions and publications needed some adjustments. In the following subsections we describe these adjustments.

12.3.1 Choice of similarity metrics

The first task to do for cleaning data was the choice of one or more metrics capable of indicating if two strings are similar or not. In the literature, several string similarity metrics have been already proposed in the past. When adopted, they are generally coupled with a threshold in such a way that two strings can be considered similar if the value returned by a similarity metric is higher than the threshold. The choice of the threshold is extremely difficult. In fact, if it is excessively low, too much false positives could be obtained; in this case, dissimilar strings would be considered similar. By contrast, an excessively high threshold would lead to too much false negatives.

In the literature the most used similarity metrics are Levenshtein, Needleman-Wunch, Smith-Waterman, Jaro, QGram Distance, Block Distance, and Jaccard Similarity. After a first analysis of the strengths and the weaknesses of the most known metrics, it was necessary to determine the most suited to our scenario. For this purpose, we considered all the institutions of a country (i.e., Afghanistan) and we applied all metrics to them. We chose Afghanistan because it was the first country in our list and because the number of its institutions made it possible a manual (and, therefore, much more precise) check of the results obtained by applying all candidate metrics.

We almost immediately determined that only one metric was not sufficient to obtain accurate results. After several tests, we found that it was sufficient to detect a pair of (at least partially) complementary metrics. Further tests showed that the most promising pair was formed by Jaccard Similarity and QGram Distance. As previously pointed out, the choice of the metrics was strictly connected with computing the most suited thresholds. For this purpose, we conducted an experimental campaign by executing an optimization algorithm, based on a hill climbing methodology. This algorithm aimed at maximizing the number of corrected results on Afghanistan data. It found that the best threshold value was 0.71 for Jaccard Similarity and 0.75 for QGram Distance.

12.3.2 Description of the algorithm for determining string similarity

After having chosen metrics and thresholds, we had to define a cleaning algorithm to use in the next steps of our ETL activity. This task was difficult. In fact, it was necessary to guarantee a possible "transitive closure" of similarities, assuming that the choice of thresholds in the previous step was capable of avoiding that an excessive usage of this closure would have led to consider as similar some strings that actually were dissimilar. To better explain this problem, consider the following example. We have three strings, namely "Paolo Russo", "Pao Russo" and "Pao Ru", representing the (possibly abbreviated) surname and name of an author. If our algorithm would have determined a similarity between "Paolo Russo" and "Pao Russo" and another similarity between "Pao Russo" and "Pao Ru", it should have been capable of understanding that there exists a similarity between "Paolo Russo" and "Pao Ru".

In order to handle transitive closure to the best, we adopted the support data structure that appeared the most adequate to conceptually representing and explaining this phenomenon, i.e., a graph. This graph G_{Sym} consists of a set N_{Sym} of nodes and a set E_{Sym} of edges. There is a node n_i for each string to evaluate. There is an edge (n_i, n_j) if the strings associated with n_i and n_j have been found to be similar by applying the Jaccard Similarity with a threshold of 0.71 and/or the QGram Distance with a threshold of 0.75.

Once G_{Sym} has been created, finding all the possible similarities among sets of strings can be carried out by finding all possible connected components in G_{Sym}. After the sets of similar strings have been found, ETL represents all the strings associated with the same university by means of a unique string in the underlying database.

12.3.3 Application of our ETL algorithm on available data

Our ETL activities concern the fields `City` and `Inst_name` of the part *Institutions*. As for `City`, our approach clusters the values of this field on the basis of the corresponding country. In this way, it avoids homonymies concerning cities having the same name but belonging to different countries. For the same reason, the values of `Inst_name` are clustered on the basis of the country, the city and the category.

For each cluster of `City`, we constructed a graph G_{Sym} and applied the algorithm described in Section 12.3.2. This algorithm computed the connected components and, for each of them, selected a city name to represent it and stored this name in `City` accordingly. For each cluster of `Inst_name`, we proceeded analogously to `City`, but the strings representing connected components were suitably stored in the field `Institution_Name1`. Furthermore, for each connected component, we registered an auto-increment number in the field `Inst_ID`.

12.4 Description of our approach

As pointed out in the Introduction, our approach is aimed to extract knowledge patterns about research activities and hubs in a set of countries of interest starting from the publications of their research institutions, as stored in the Web of Science repository. Before starting its description, we must define some sets that formalize available data and, therefore, will be extensively used below.

The first set regards the set RA of research areas. It consists of the following elements:

$$RA = \{\text{‘NS’, ‘AS’, ‘MH’, ‘SS’, ‘HU’, ‘ET’}\}$$

where ‘NS’ (resp., ‘AS’, ‘MH’, ‘SS’, ‘HU’, ‘ET’) stands for ‘Natural Science’ (resp., ‘Agricultural Science’, ‘Medical and Health Science’, ‘Social Science’, ‘Humanities’, ‘Engineering and Technology’).

The second set concerns the overall set Pub of publications at our disposal. Given a publication $p \in Pub$, we indicate by $Authors_p$ the set of its authors and by $Areas_p$ the set of the research areas it belongs to.

The third basic set regards the set C of the countries to investigate.

12.4.1 Hub characterization and detection

In this section, we define a method for detecting both hubs and their features in a set of countries. For this purpose, we preliminarily introduce a first support data structure. It is a social network:

$$G = \langle N, E \rangle$$

N is the set of the nodes of G. A node $n_i \in N$ corresponds to exactly one institution registered in our database. Since there is a biunivocal correspondence between a node of N and the corresponding institution, in the following, we will use the symbol n_i to indicate both of them. Each node of N is labeled with an element of C depending on the country of the corresponding institution. We indicate by l_i the label of n_i. E is the set of the edges of G. There exists an edge $e_{ij} = (n_i, n_j, w_{ij}) \in E$ if there exists at least one publication involving one author of n_i and one author of n_j. w_{ij} is the weight of e_{ij}; it denotes the number of publications having at least one researcher of n_i and one researcher of n_j among their authors.

Starting from this support structure, we can now define some sets regarding the neighborhoods of a node in G. Specifically, we define the neighborhood nbh_i of a node $n_i \in N$ as the set of the nodes of G directly connected with n_i:

$$nbh_i = \{n_j | (n_i, n_j, w_{ij}) \in E, n_j \neq n_i\}$$

Then, we can define the sets nbh_i^I (resp., nbh_i^F) of the neighbors of n_i belonging to the same country as (resp., to different countries from) the one of n_i:

$$nbh_i^I = \{n_j | n_j \in nbh_i, l_i = l_j\} \qquad nbh_i^F = \{n_j | n_j \in nbh_i, l_i \neq l_j\}$$

Now, we introduce the set N_k of the nodes (i.e., institutions) of a given country k:

$$N_k = \{n_i | n_i \in N, l_i = k\}$$

Another group of sets can be defined for representing several features about publications:

$$Pub_k = \{p \in Pub | \text{ at least one element of } Authors_p \text{ operates at an institution of } N_k\}$$
$$Pub_k^I = \{p \in Pub | \text{ all the elements of } Authors_p \text{ operate at institutions of } N_k\}$$

After this, we introduce Pub_{ij} as the set of publications simultaneously having researchers of both n_i and n_j as their authors. We also introduce the set $JPub$ (resp., $CPub$) as the set of papers published in a journal (resp., proceedings of a conference). We define $JCPub$ as $JCPub = JPub \cup CPub$. We also define the set Pub^q of the publications belonging to the research area q:

$$Pub^q = \{p \in Pub | q \in Areas_p\}$$

Finally, we define Pub_k^q as the subset of Pub^q having at least one author operating at an institution of the country k.

Now, we are able to introduce the concept of hub. For this purpose we need to introduce three metrics.

The first metric, M_1, is defined in such a way that, given a node n_i, M_{1_i} is equal to the sum of the weights of the edges linking n_i. Observe that this metric coincides with the classical weighted degree centrality [193, 460, 246, 238, 9]. Formally speaking:

$$M_{1_i} = \sum_{j \in nbh_i} w_{ij}$$

The second metric, M_2, is defined in such a way that, given a node n_i, M_{2_i} is the ratio of the sum of the weights of the edges linking n_i to nodes associated with foreign institutions to the average number of publications relative to the country of n_i. Observe that this metric coincides with the normalized weighted degree centrality [193, 460, 246, 238, 9]. Formally speaking:

$$M_{2_i} = \frac{\sum_{j \in nbh_i^F} w_{ij}}{AvgPub_k}$$

where $AvgPub_k = \frac{\sum_{n_i \in N_k, n_j \in N, n_i \neq n_j} w_{ij}}{|N_k|}$.

The third metric, M_3, is analogous to M_2 except that, in the numerator, the sum of the weights of the edges linking n_i to nodes of the same country is considered, since this metric takes publications with internal institutions into account. Interestingly, this metric is analogous to the E-I index [193, 304, 199, 435]. Formally speaking:

$$M_{3_i} = \frac{\sum_{j \in nbh_i^I} w_{ij}}{AvgPub_k}$$

According to both the theoretical and the experimental results described in [193, 460, 246, 238, 9, 304, 199, 435], and as verified in our case study (see Section 12.5.1), M_1, M_2 and M_3 follow a power law distribution.

Taking all these considerations into account, the set $\mathcal{H}^X$ of hubs can be defined as the set of those institutions simultaneously belonging to the top $X\%$ of the institutions with the highest values of M_1, M_2 and M_3 (we call I_1^X, I_2^X and I_3^X these three sets, when considered separately).

The set $\mathcal{H}^X$ of hubs is defined as:

$$\mathcal{H}^X = \{n_i \in N | n_i \in (I_1^X \cap I_2^X \cap I_3^X)\}$$

where:

$$I_1^X = \{n_i \in$$

$N | M_{1_i}$ belongs to the top $X\%$ of the values of M_1, when applied to the nodes of $N\}$

$$I_2^X = \{n_i \in$$
$N|M_{2_i}$ belongs to the top $X\%$ of the values of M_2, when applied to the nodes of $N\}$
$$I_3^X = \{n_i \in$$
$N|M_{3_i}$ belongs to the top $X\%$ of the values of M_3, when applied to the nodes of $N\}$

In this definition, X is a threshold allowing the selection of the institutions having the highest values of M_1, M_2 and M_3. The choice to use X as a threshold parameter derives from the power law distributions characterizing all the three metrics. Reasonable values of X could be 10, 15 and 20. After several experiments (see Section 12.5.1), we decided to consider a default value of X equal to 20. As a consequence, in the following, when X is not specified, we intend that it is equal to 20.

The rationale underlying this definition is that a hub is an institution that simultaneously belongs:

- to the top $X\%$ of the institutions publishing more papers (we call this condition C_1; it is handled by metrics M_1);
- to the top $X\%$ of the institutions publishing more papers with institutions of a country different from their own (we call this condition C_2; it is handled by metrics M_2);
- to the top $X\%$ of the institutions publishing more papers with institutions of their own country (we call this condition C_3; it is handled by metrics M_3).

It is worth pointing out that our hub definition could be seen as an attempt to introduce a new form of node centrality (specific to the context of interest), which takes into account both the number of edges relative to a node and their weights. In this sense, our hub definition follows the same general philosophy proposed in [340], where the authors present new versions of degree, closeness and betweenness centrality that take both incoming edges and their weights into consideration.

In the following, we use the symbol $\mathcal{H}_k^X$ to indicate the hubs of a given country k. The application of the parameters introduced in this section to our case study can be found in Section 12.5.1.

12.4.2 Investigation of the research scenarios for the countries of interest

In this section, we aim at analyzing the research scenarios of the countries of C in such a way as to detect their most important features and to highlight similarities and differences among them. Initially, we define I_1' as the set of the institutions of I_1 belonging to a country of C.

Now, we can introduce three indicators that could give us some knowledge about the research scenarios of the countries of C.

- The first one, RQ, is an indicator of the overall research quality in the countries of C:

$$RQ = \frac{|I'_1|}{|I_1|}$$

- The second one, FC, indicates how many institutions, among the top ones of the countries of C, publish many papers with foreign institutions:

$$FC = \frac{|I'_1 \cap I_2|}{|I'_1|}$$

- The third one, TP, indicates how many institutions that publish very much with foreign institutions belong to the top institutions of the countries of C:

$$TP = \frac{|I'_1 \cap I_2|}{|I_2|}$$

In the investigation of the research scenario of a country k and of the role of its hubs, it appears very interesting to study its paper distribution. For this purpose, we introduce the average number $AvgPub_k^{\mathcal{H}}$ of the publications of its hubs:

$$AvgPub_k^{\mathcal{H}} = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N, n_i \neq n_j} w_{ij}}{|\mathcal{H}_k|}$$

Another interesting issue to investigate is to verify if a hub of k publishes more with institutions of k (we call "internal" the corresponding publications) than with foreign ones (we call "external" the corresponding publications) or alone. To carry out this investigation, we introduce:

- the average number $AvgHubPub_k^I$ of publications performed by the hubs of $\mathcal{H}_k$ with other institutions of the same country (here, the apex "I" stands for "Internal"):

$$AvgHubPub_k^I = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N_k, n_i \neq n_j} w_{ij}}{|\mathcal{H}_k|}$$

- the average number $AvgHubPub_k^F$ of publications performed by the hubs of $\mathcal{H}_k$ with other institutions of a country different from k (here, the apex "F" stands for "Foreign"):

$$AvgHubPub_k^F = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N - N_k} w_{ij}}{|\mathcal{H}_k|}$$

- the average number $AvgHubPub_k^A$ of publications performed alone by the hubs of $\mathcal{H}_k$, i.e., with authors that belong all to the same institution of the country k (here, the apex "A" stands for "Alone"):

$$AvgHubPub_k^A = \frac{\sum_{n_i \in \mathcal{H}_k} w_{ij}}{|\mathcal{H}_k|}$$

A further interesting analysis is devoted to understand if, in their cooperation with foreign institutions, the hubs of $\mathcal{H}_k$ privilege one or few countries. For this purpose, we specialize the Herfindahl Index to our research context. Specifically, in our case, we define the Herfindahl Index HI_k associated with the papers published by the hubs of $\mathcal{H}_k$ to verify if these hubs published in cooperation with institutions of few (implying high values of HI_k) or many (implying low values of HI_k) countries.

In order to apply the Herfindahl index to our context, we must introduce the following support parameters:

- number of publications that the hubs of $\mathcal{H}_k$ performed with foreign institutions: $PubH_k^F = \sum_{n_i \in \mathcal{H}_k, n_j \in N - N_k} w_{ij}$
- fraction of the external publications that the hubs of $\mathcal{H}_k$ performed with the institutions of a country q: $PubFr_{kq}^F = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N_q} w_{ij}}{PubH_k^F}$
- set of countries having at least one paper with the institutions of a country k: $Cntr_k^F = \{q | \exists (n_i, n_j, w_{ij}) \in E, n_i \in \mathcal{H}_k, n_j \in N_q\}$

We can, now, define the Herfindahl Index associated with the papers published by the hubs of $\mathcal{H}_k$ as follows:

$$HI_k = \sum_{q=1..|Cntr_k^F|} \left(PubFr_{kq}^F \right)^2$$

The possible values of the Herfindahl Index range in the real interval $\left[\frac{1}{|Cntr_k^F|}, 1 \right]$, where $\frac{1}{|Cntr_k^F|}$ is obtained when each paper is published with an institution of a different country, and 1 in the opposite case.

The application of the parameters introduced in this section to our case study can be found in Section 12.5.2.

Cooperation among hubs of the same country

In this section, we aim at investigating the cooperation levels of the hubs $\mathcal{H}_k$ of a given country k. For this purpose, we preliminarily define a support data structure called *clique social network*.

In particular, let G be the social network defined in Section 12.4.1 and let G_k be its "projection" on the country k. Let $\mathcal{C}_k$ be the set of cliques of G_k and let $\mathcal{H}_k$ be the set of the hubs of k. A *clique social network* CG_k has a node for each hub of $\mathcal{H}_k$ belonging to at least one clique of $\mathcal{C}_k$. Each node n_i of CG_k has associated a weight w_i denoting the number of cliques of $\mathcal{C}_k$ which it belongs to. An edge (n_i, n_j) of CG_k denotes that n_i and n_j together belong to at least one clique of $\mathcal{C}_k$.

Some measures capable of quantitatively representing the differences that characterize the cooperation among hubs are the following: *(i)* the number of cliques $|\mathcal{C}_k|$; *(ii)* the absolute dimension $d_{\mathcal{C}_k}$ of the largest clique in $\mathcal{C}_k$; *(iii)* the relative dimension

$\frac{d_{\mathcal{C}_k}}{|\mathcal{H}_k|}$ of the largest clique in $\mathcal{C}_k$; *(iv)* the fraction $f^{\mathcal{H}}_{\mathcal{C}_k}$ of hubs belonging to at least one clique of $\mathcal{C}_k$.

In order to avoid that results are biased by the number of publications (which can be very different in the different countries of interest), we define a normalized version $\widehat{CG_k}$ of CG_k.

$\widehat{CG_k}$ is obtained by performing the same steps carried out for constructing CG_k but on a graph $\widehat{G_k}$, instead of on G_k. $\widehat{G_k}$ has the same nodes as G_k. There is an edge (n_i, n_j) in $\widehat{G_k}$ only if $\frac{Pub_{ij}}{Pub_k}$ is higher than a threshold th. We have experimentally verified that, generally, $\frac{Pub_{ij}}{Pub_k}$ follows a power law distribution. As a consequence, we have chosen to set th in such a way as to discard the 20% of the lowest values of $\frac{Pub_{ij}}{Pub_k}$.

Finally, we searched for some measures to compare clique social networks. After several experiments, we found that the most significant ones were: *(i)* the number of nodes; *(ii)* the number of edges; *(iii)* density[1].

The application of these parameters to our case study is reported in Section 12.5.2.

12.4.3 Investigation of research areas

All reasonings and computations performed above for countries can be repeated for research areas. For this purpose, we define a support social network, called *RA social network*. In particular, the *RA social network* S_q, associated with the research area q, is defined as:

$$S_q = \langle N_q, E_q \rangle$$

Here, N_q is the set of the nodes of G having at least one publication belonging to Pub^q. There exists an edge $e_{ij} = (n_i, n_j, w_{ij}) \in E_q$ if there exists at least one publication of Pub^q involving one author of n_i and one author of n_j. w_{ij} indicates the number of publications of Pub^q performed by at least one author of n_i and an author of n_j.

Another important parameter, very useful in this context, is the set $\mathcal{H}_q$ of the hubs related to the research area q. Its detailed definition and the way to compute it are analogous to the corresponding ones we have described for $\mathcal{H}$ in Section 12.4.1.

The application of these data structures and concepts to our case study can be found in Section 12.5.3.

12.4.4 Investigation of the quality of publications

All indicators introduced above are based only on the number of publications. However, it would be important to take also their quality into account. One way to do this

[1] As a matter of facts, this last measure can be derived from the two other ones, but it is very expressing and, consequently, we decided to explicitly consider it.

consists in taking impact factor into consideration; another way consists in considering the number of citations received by papers.

Impact factors are measured only for journal papers. As a consequence, if we want to employ this measure, we must define a new support data structure. This structure, that we indicate by G', is, once again, a social network. It is defined as:

$$G' = \langle N', E' \rangle$$

There is a node $n_i \in N'$ for each institution having at least one author that published at least one journal paper. An edge $e'_{ij} = (n'_i, n'_j, w'_{ij})$ has a semantics similar to the one of e_{ij} except that the weight $w'_{ij} = \sum_{p \in (Pub_{ij} \cap JPub)} IF_p$ considers both the number of publications simultaneously performed by n_i and n_j and the corresponding impact factors.

Paper citations are valid both for conference proceedings and for journal papers. However, in order to make our analyses about the quality of publications homogeneous, we chose to investigate only journal papers. In this case, we used the same support social network as the one employed for impact factors but the edge weights w'_{ij} was computed as: $w'_{ij} = \sum_{p \in (Pub_{ij} \cap JPub)} CitN_p$, where $CitN_p$ is the number of citations of p.

The application of these data structures and concepts to our case study is reported in Section 12.5.4.

12.4.5 Characterization of hub neighborhoods

A first parameter useful to characterize hub neighbors is the average number $AvgPub$ of publications of the hub neighborhoods. It is defined as:

$$AvgPub = \frac{\sum_{i \in \mathcal{H}} AvgNbhPub_i}{|\mathcal{H}|}$$

where $AvgNbhPub_i = \frac{\sum_{n_j \in nbh_i^I} \sum_{n_k \in nbh_j^I} w_{jk}}{|nbh_i^I|}$.

Since, in the hub neighborhoods, there could be other hubs, which clearly can strongly influence the neighborhood behavior, we define an additional version of hub neighborhoods $\widehat{nbh_i}$, $\widehat{nbh_i^I}$ and $\widehat{nbh_i^F}$, obtained by filtering out hubs from nbh_i, nbh_i^I and nbh_i^F, respectively. Then, we define $\widehat{AvgPub}$ by simply substituting nbh_i with $\widehat{nbh_i}$.

We call $AvgPub_k$ (resp., $\widehat{AvgPub}_k$) the "projection" of $AvgPub$ (resp., $\widehat{AvgPub}$) on the country k: $AvgPub_k = \frac{\sum_{i \in \mathcal{H}_k} AvgNbhPub_i}{|\mathcal{H}_k|}$.

A second parameter for evaluating hub neighborhoods regards their average dimension $AvgDim$:

$$AvgDim = \frac{\sum_{i \in \mathcal{H}} |nbh_i|}{|\mathcal{H}|}$$

Also in this case, we disaggregate data per country and we call $AvgDim_k$ the corresponding parameter for the country k.

A next analysis regards the cooperation level among the institutions belonging to neighborhoods. To perform this task, we define a new support social network. We call it *nbh social network* and we represent it by means of the symbol $NbhG_i$. Given a neighborhood nbh_i, the corresponding nbh social network is defined as follows:

$$nbhG_i = \langle nbh_i, nbhE_i \rangle$$

There is a node in $NbhG_i$ for each node of nbh_i; there is an edge $(n_i, n_j) \in nbhE_i$ if there exists at least one publication between an author of n_i and an author of n_j.

After having introduced this social network, we define a first parameter on it. This parameter is called $AvgCFrac$ and corresponds to the average fraction of the real number of cliques existing in hub neighborhoods against the possible number of them. It is an indicator of the cooperation level among hubs. It is defined as:

$$AvgCFrac = \frac{\sum_{i \in \mathcal{H}} NbhCFrac_i}{|\mathcal{H}|}$$

Here, $NbhCFrac_i = \frac{\widetilde{C_i}}{2^{|nbh_i|} - |nbh_i| - \frac{|nbh_i|(|nbh_i|-1)}{2}}$, where $\widetilde{C_i}$ represents the number of cliques in $NbhG_i$, whereas the denominator of $NbhCFrac_i$ indicates the possible number of cliques in nbh_i. As usual, we call $AvgCFrac_k$ the "projection" of $AvgCFrac$ on the country k.

A second parameter about intra-neighborhood cooperation regards the average fraction $AvgCNbh$ of the number of cliques existing in hub neighborhoods against the number of neighborhood nodes:

$$AvgCNbh = \frac{\sum_{i \in \mathcal{H}} NbhCNum_i}{|\mathcal{H}|}$$

Here, $NbhCNum_i = \frac{\widetilde{C_i}}{|nbh_i|}$. Again, we call $AvgCNbh_k$ the "projection" of $AvgCNbh$ on the country k.

A final parameter measuring the cooperation level between hub neighbors is the average density $AvgDens$ of the nbh social network:

$$AvgDens = \frac{\sum_{i \in \mathcal{H}} NbhSNDens_i}{|\mathcal{H}|}$$

Here, $NbhSNDens_i = \frac{|nbhE_i|}{\frac{|nbh_i|(|nbh_i|-1)}{2}}$. As usual, we call $AvgDens_k$ the "projection" of $AvgDens$ on the country k.

The application of these parameters to our case study is reported in Section 12.5.5.

12.5 Application of our approach to four North African countries

As pointed out in the Introduction, we applied our approach to four North African countries, namely Algeria, Egypt, Morocco and Tunisia. As a consequence, in our case study, the set C introduced in Section 12.4, consisted of the following elements:

$$C = \{\text{‘A’, ‘E’, ‘M’, ‘T’, ‘O’}\}$$

where ‘A’ (resp., ‘E’, ‘M’, ‘T’, ‘O’) stands for ‘Algeria’ (resp., ‘Egypt’, ‘Morocco’, ‘Tunisia’, ‘Others’). Clearly, ‘O’ does not represent a specific country, but it indicates all the ones different from the four into examination. The reasons for adding ‘O’ will be clear below.

Our dataset was stored in a MongoDB database [6]. To give an idea of it, we report some of its features: *(i)* dimension = 10.27 GB; *(ii)* number of institutions = 278,696 ; *(iii)* number of authorships = 89,008,846; *(iv)* number of publications = 6,599,104; *(v)* number of research areas = 6; *(vi)* number of research fields = 251.

12.5.1 Hub characterization and detection

We computed the distribution of M_1 for the publications of $JCPub$ for each year. For instance, in Figure 12.1, we show the distribution of M_1 for $JCPub$ in the year 2013. It is a very steep power law distribution; in other cases, the trends are less steep, but, anyhow, they follow power law distributions. We do not report the other trends for space reasons; in any case, all of them are similar to the one of Figure 12.1. Obtained results confirm that, in our case study, the theoretical conjecture about the trend of weighted degree centrality [193] is valid.

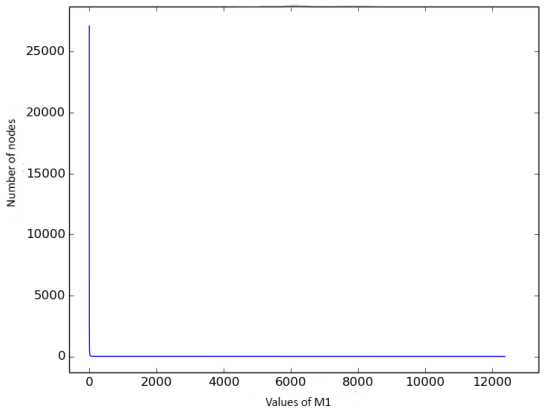


Fig. 12.1. Distribution of M_1 for the publications of $JCPub$ in the year 2013

Then, we computed the distribution of M_2 and M_3 for the publications of $JCPub$ for each year. Analogously to what happened for M_1, all the trends were the same and followed power law distributions, thus confirming what theoretically said in [193].

In order to understand the filtering level of hubs against the increase of X, and in order to choose a default value for this parameter, we computed the number of hubs belonging to the four countries into consideration over time for the three values of X chosen in Section 12.4.1. We report obtained results in Figure 12.2. From the analysis of this figure, we can see that the trend of selected hubs is always increasing over time and very similar for the three values of X. This implies that all the three values of X would lead to the same behavior of our approach. The only difference regards the desired tradeoff between the number and the strength of the identified hubs. The higher X, the stronger (but, the less numerous) the identified hubs. We have preferred to let our approach to be more "permissive", i.e., to let it privilege hub number on hub strength. As a consequence, we set X to a default value of 20. However, in case of hub strength needs be privileged on hub number, it would be sufficient to set X to a low value, for instance to set it to 10.

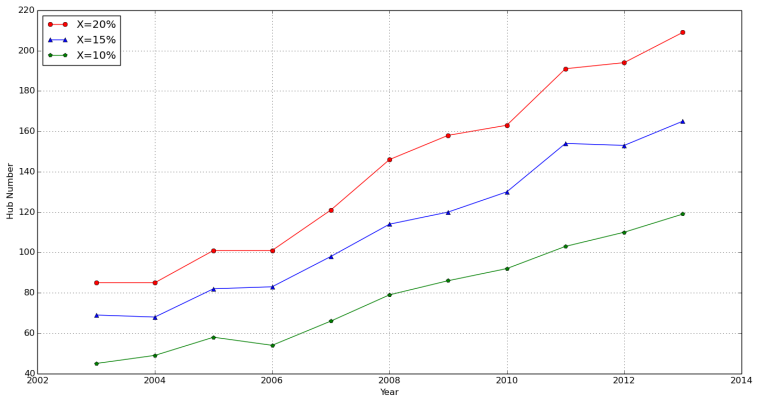


Fig. 12.2. Hub number over time for several values of X

In many research fields, conferences are not considered in the computation of bibliometric indices. As a consequence, we judged interesting to remake all the previous investigations considering journals only. This corresponded to analyze publications belonging to $JPub$, instead of to $JCPub$. All the analyses performed for $JPub$ confirmed the general trends and the results found for $JCPub$. For instance, also in this case, M_1, M_2 and M_3 presented a power law distribution for all the four countries into consideration. Interestingly, in case of $JPub$, power law distributions are generally steeper than the ones of $JCPub$.

Another very interesting, and quite unexpected, result regards the number of hubs when only journals are considered. In fact, although the number of involved institutions decreases, the number of hubs generally does not decrease and, in several cases, increases. This quite surprisingly result can be explained by considering that the publication of a paper on conference proceedings has quite high costs (think, for instance, of costs for conference registration, travel, stay, etc.). These can disadvantage the institutions of the four countries into consideration, since all of them are characterized by a low average income per capita. If we consider $X = 10$ or $X = 15$, we obtain the same results.

An important characterization of hubs regards their capability of cooperating each other. In other words, it is interesting to verify if there exists a sort of backbone comprising hubs of different countries. To perform this investigation, we considered the concept of clique. Recall that a clique of dimension η is simply a complete subgraph consisting of η nodes. To conduct our analysis we carried out the following steps:

- We considered two time intervals. The former is $[2003, 2009]$, the latter is $[2007, 2013]$. We considered them expressly overlapped to avoid the risk of discontinuity.
- We "projected" the social network G in two social networks G' and G'' in such a way as to consider only hubs and only publications of the period $[2003, 2009]$ in the former, and of the period $[2007, 2013]$ in the latter.
- We computed all the cliques of G' and G''.

After this, we analyzed the number and the dimension of obtained cliques, as well as the institutions belonging to them. As a general trend, we found that there are many cliques and most of them are very small. This indicates that there are some contacts among hubs but there is not a strict cooperation among many of them in such a way as to have "research backbones".

Furthermore, the largest clique of the period $[2003, 2009]$ consisted of 13 hubs, whereas the largest one of the period $[2007, 2013]$ was formed by 17 hubs. In both cases all hubs forming these cliques are only Egyptians. From this analysis, we can draw the following knowledge patterns:

- Cliques tend to enlarge over time, although slowly. For instance, the largest clique of the period $[2007, 2013]$ is obtained by aggregating four further hubs to those belonging to the largest clique of the period $[2003, 2009]$.
- The largest cliques are formed by hubs of the same country; for instance, the top 5 cliques in the two periods are all formed by Egyptian hubs only. This last result has a further important consequence in that it shows that hubs of different countries tend to not cooperate each other.

Years	RQ	FC	TP
2003	0.0593	0.667	0.785
2004	0.0572	0.731	0.819
2005	0.0577	0.748	0.865
2006	0.0598	0.616	0.850
2007	0.0574	0.638	0.860
2008	0.0555	0.629	0.830
2009	0.0612	0.602	0.852
2010	0.0555	0.621	0.891
2011	0.0516	0.658	0.892
2012	0.0503	0.660	0.888
2013	0.0471	0.701	0.894

Table 12.1. Values of RQ, FC, and TP in the year interval [2003,2013] when both conferences and journals are considered

12.5.2 Investigation of the research scenarios for the countries of interest

First of all, we computed the three indicators RQ, FC and TP, whose formalization has been provided in Section 12.4.2, for the four North African countries of interest. This computation (for both $JCPub$ and $JPub$) returned very interesting knowledge patterns about the research scenarios in the four countries (see Table 12.1). In particular, the first indicator shows that the research institutions in the four countries do not present excellent performances. Furthermore, this indicator does not show a significant increase over time. The second and the third indicators highlight that an institution of one of the four countries benefits very much from the cooperation with foreign institutions for reaching and maintaining a high performance in its own country.

After these analyses, we started to investigate the similarities and the differences for hubs in the four countries. First, we computed the values of M_1, M_2 and M_3 in the four countries for all the years into consideration. We obtained that both M_1 an M_3 present a power law distribution and, therefore, confirm what we have seen for the general case. An interesting trend is shown by M_2 for these countries (Figure 12.3). Indeed, this measure presents a distribution characterized by a broken line with quite a rapid decrease and a possible starting peak. This suggests a very interesting scenario for the hubs in each of the four countries. This scenario is the typical one of an oligarchy of hubs for each country and is very different from the two ones we had initially hypothesized (i.e., a lot of quite weak hubs, corresponding to a smoothly decreasing distribution for M_2, or a very few number of very strong hubs, corresponding to a power law distribution for M_2).

In Figure 12.4, we report the variation of the number of hubs for each country. From the analysis of this figure, we can see that the country with the highest number of hubs is Tunisia. This result was unexpected also because both the extension and the number of citizens of Tunisia were smaller than the ones of the other three countries.

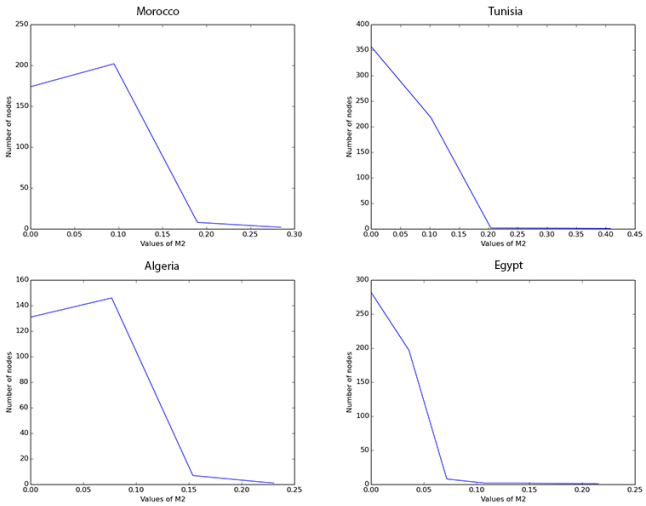


Fig. 12.3. Trend of M_2 for the four countries in the year 2013

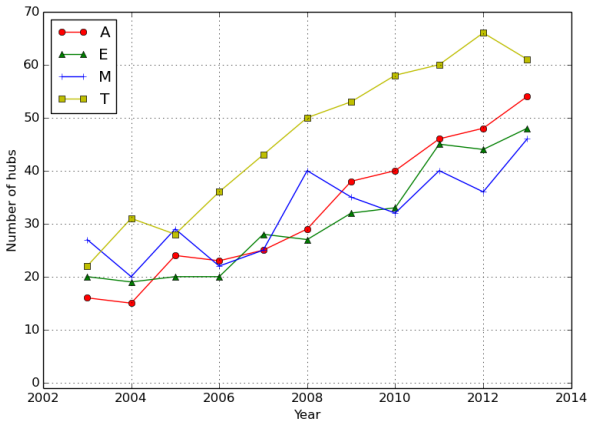


Fig. 12.4. Number of hubs for each country in the year interval [2003,2013]

In Figure 12.5, we report the values of the average number $AvgPub_k^{\mathcal{H}}$ of hub publications over time (see Section 12.4.2) for the four countries of interest. From the analysis of this figure we can see that Egyptian hubs generally publish much more papers than the hubs of the other countries. This result, along with the ones of Figure 12.4, suggests that research in Egypt is much more concentrated than in the other three countries.

A final report about this issue regards the total number of publications $|Pub_k|$ (see Section 12.4.1) over time for each country. Obtained results evidence that that Egypt has a number of publications much higher than the other three countries. This result, along with the ones reported in Figure 12.4, is a further confirmation that research in Egypt is much more concentrated than in the other three countries.

After this, we computed the average number of internal, external and alone publications for the four countries of interest. Obtained results evidence that the hubs of

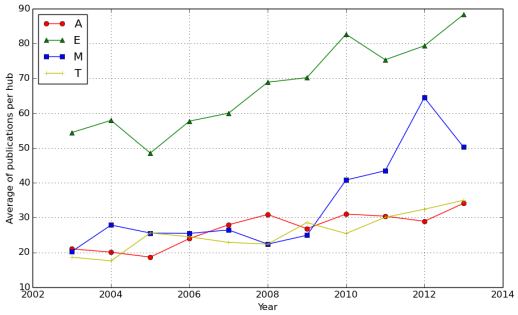


Fig. 12.5. Average number of publications per hub over time for the four countries

all countries always publish more with foreign institutions than with internal ones. Interestingly, Egyptian hubs have a significant fraction of alone publications.

In Figure 12.6, we report the Herfindahl index HI_k for the four countries (see Section 12.4.2). From the analysis of this figure we can observe that Tunisia and Algeria have a high Herfindahl index, which implies that their hubs cooperate mostly with one or few countries. By contrast, Egypt has a very low Herfindahl index, i.e., its hubs cooperate with many countries. An interesting trend is the one of Morocco; in fact, it initially has a behavior like the ones of Tunisia and Algeria, whereas, in the last years, it shows a behavior like the one of Egypt.

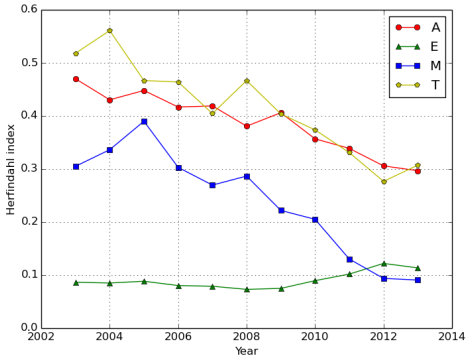


Fig. 12.6. Herfindahl index over time for the four countries

A possible objection to the previous way of proceeding could be that the computation of the Herfindahl index of a country k (e.g., Egypt) could be "biased" by the presence of many institutions of different countries each having only one publication with a hub of k. To overcome this objection, for each country k, we considered the top 5 countries T_k sharing publications with its hubs. Then, we recomputed the Herfindahl index considering, for each k, only the institutions belonging to the countries of T_k. Obtained results show that all the main conclusions we have drawn from Figure 12.6 are still valid. This not only overcomes the previous objection, but it is also a

further confirmation of the power law distribution of hubs' publications, which we have detected by studying the trend of M_1.

Cooperation among hubs of the same country

To determine the cooperation levels among hubs for the four North African countries into consideration, for each country k, we performed the following tasks:

- We considered the two time intervals $[2003, 2009]$ and $[2007, 2013]$.
- We computed the clique social networks (see Section 12.4.2) $CG1_k$ (resp., $CG2_k$), corresponding to the first (resp., the second) time interval.
- We measured the four parameters introduced in Section 12.4.2 for quantitatively evaluating clique social networks.

Obtained results are reported in Table 12.2. From the analysis of these tables we can draw the following conclusions:

- Egypt has the largest clique in both periods; the clique is much larger than the maximum cliques of the other countries;
- in Egypt almost all hubs belong to at least one clique.

Country	$\lvert\mathcal{C}1_k\rvert$	$d1_{\mathcal{C}_k}$	$\frac{d1_{\mathcal{C}_k}}{\lvert\mathcal{H}_k\rvert}$	$f1^{\mathcal{H}}_{\mathcal{C}_k}$
Algeria	292	7	0.152	0.913
Egypt	38	13	0.351	0.973
Tunisia	130	8	0.116	0.942
Morocco	82	7	0.127	0.818

Country	$\lvert\mathcal{C}2_k\rvert$	$d2_{\mathcal{C}_k}$	$\frac{d2_{\mathcal{C}_k}}{\lvert\mathcal{H}_k\rvert}$	$f2^{\mathcal{H}}_{\mathcal{C}_k}$
Algeria	234	8	0.121	0.939
Egypt	94	17	0.27	1.0
Tunisia	304	9	0.081	0.847
Morocco	106	9	0.134	0.821

Table 12.2. Quantitative differences characterizing the cooperation behaviors of hubs in the four countries (first time interval on the top and second time interval on the bottom)

These results indicate that Egyptian hubs are more prone to cooperation than the hubs of the other countries.

In Figure 12.7, we report the graphs $CG2_k$ for all the four countries; in these graphs the dimension of nodes is proportional to the corresponding weight, i.e., to the number of cliques they belong to. The analysis of this figure confirms the previous conjecture; in fact, the number of edges in the Egyptian graph is much higher than in the other graphs. This fact, along with the presence of many little nodes, allows us to derive another important knowledge pattern, i.e., that research cooperation in Egypt is more advanced than in the other countries.

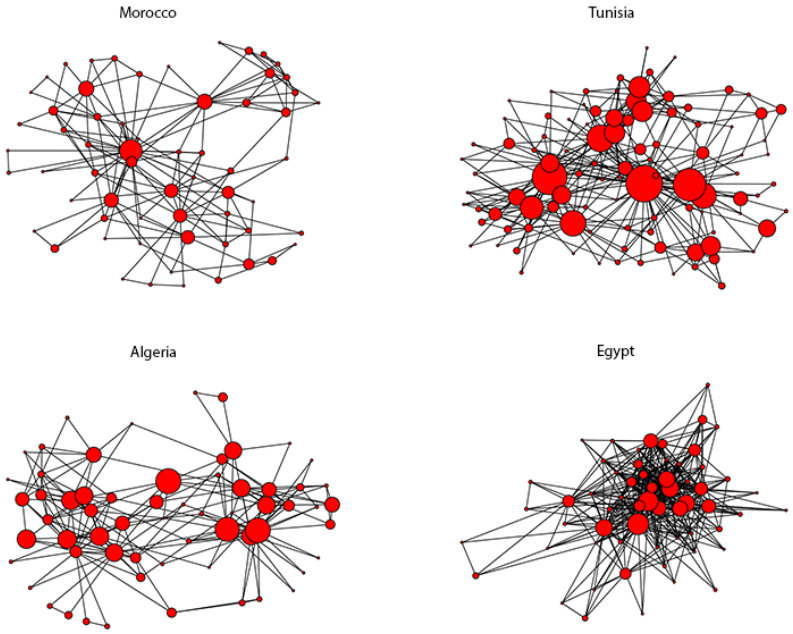


Fig. 12.7. Graphs $CG2_k$ for the four countries

In Figures 12.8 and 12.9, we report the graphs $\widehat{CG1_k}$ and $\widehat{CG2_k}$ (corresponding to $\widehat{CG_k}$ for the first and the second time interval) for the four countries. From the analysis of these figures we can observe that the different behavior of Egyptian hubs with respect to the ones of the other countries is confirmed, although slightly attenuated.

Finally, we computed the number of nodes, the number of edges and the density of $CG1_k$ and $CG2_k$ for all countries. Obtained results are reported in Table 12.3. From the analysis of this table we can observe that the three measures quantitatively depict very well what we have expressed previously. Furthermore, if we compare their values in the two periods, we can draw some interesting knowledge patterns. In fact, we can observe that the number of nodes always increases, which implies an increase of the hub capability of cooperating each other. This increase is quite high (i.e., about 38%) for Morocco, high (i.e., about 50%) for Algeria and Tunisia, and very high (i.e., about 76%) for Egypt. The same trends can be observed for the increase of the number of edges (i.e., about 27% for Morocco, about 50% for Algeria and Tunisia, and about 129% for Egypt). By contrast density always decreases. These last results represent a further confirmation about the fact that hubs continue to cooperate a little each other.

Finally, if we consider the ratio of the increase of the number of edges to the increase of the number of nodes for the four countries when passing from the first to the second time interval, we can observe that this ratio is about 1 for Algeria and Tunisia, about 1.70 for Egypt and about 0.73 for Morocco. This indicates that, in the

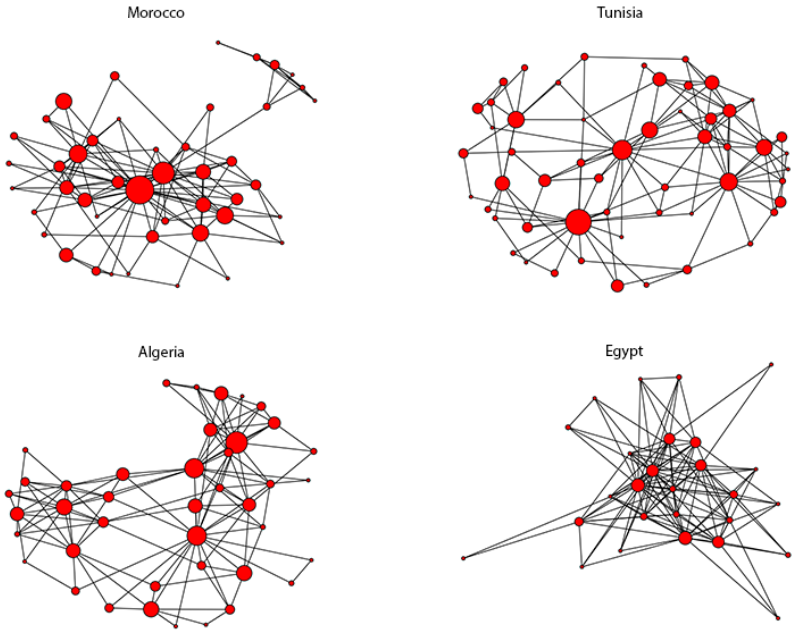


Fig. 12.8. Graphs $\widehat{CG1_k}$ for the four countries

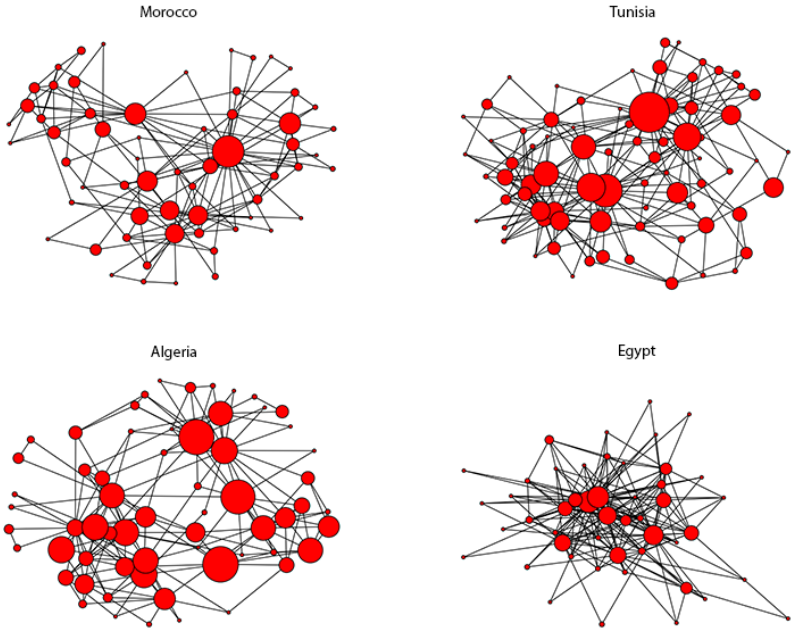


Fig. 12.9. Graphs $\widehat{CG2_k}$ for the four countries

Country	*number of nodes*	*number of edges*	*density*
Algeria	41	166	0.200
Egypt	34	196	0.349
Tunisia	66	272	0.127
Morocco	45	174	0.176

Country	*number of nodes*	*number of edges*	*density*
Algeria	62	249	0.132
Egypt	60	450	0.254
Tunisia	103	416	0.079
Morocco	62	221	0.117

Table 12.3. Number of nodes, number of edges and density of $CG1_k$ (on the top) and of $CG2_k$ (on the bottom) for all countries

second time interval, Egypt had a spectacular increase of the hub cooperation. This also reflects in the density decrease, which is much more reduced in Egypt than in the other countries (in fact, it is about -27% for Egypt, -34% for Algeria, -38% for Tunisia and -59% for Morocco). As for density, its decrease must not be misleading since, to avoid it, the number of edges should have increased against the square of the number of nodes, which is almost impossible. As a matter of fact, the number of edges always increases in all the four countries, but slightly.

12.5.3 Investigation of research areas

For each research area of *RA* (see Section 12.4), we computed the corresponding *RA* network and, then, we repeated all the tasks described in the previous sections in such a way as to disaggregate the corresponding results per research area.

A first analysis regarded the distribution of M_1, M_2 and M_3 over time for each research area. In this case, we obtained that these distributions are analogous to the ones obtained for aggregated data.

For each research area, we computed the number of publications of hubs over time. Obtained results show that the research areas having the highest number of hubs are 'NS', 'ET' and 'MH'. This result confirms the ones reported in [251] concerning the diffusion of research areas in the same countries.

Then, we computed the number of publications per hub over time for each research area. Obtained results are in line with the ones shown in the previous figures.

A further, very interesting, disaggregation of results is obtained by separating data for pairs (country, research area). In fact, in this way, we can verify if the four countries into consideration present similar or dissimilar features and behaviors in the different research areas.

A first analysis of this disaggregation level regarded the distribution of M_1, M_2 and M_3. Obtained results confirm that these metrics follow a power law distribution

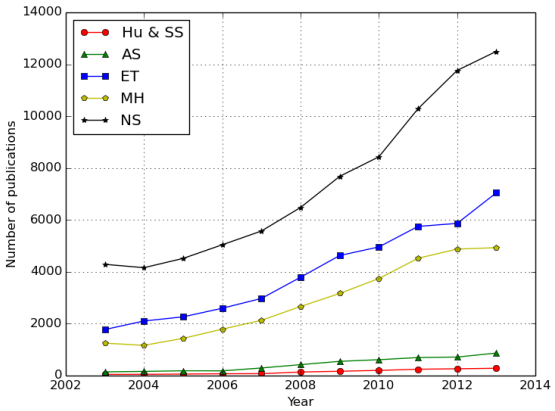


Fig. 12.10. Average number of publications of hubs over time for each research area

for 'NS', 'ET' and 'MH'. For the other three research areas, the number of publications performed in the four countries was small and, therefore, we had to discard obtained results, because they were not reliable.

A second investigation at the same disaggregation level concerned the number of hubs over time. Obtained results confirm in principle the ones about the distribution of hubs per country, shown in Figure 12.2.

We have seen that a particular feature of hubs was the fact that they published more with foreign institutions than with internal ones (see Section 12.4.2). We repeated this investigation at the new disaggregation level and found that this trend is always valid except for the pair (Tunisia, 'MH'), where it is never valid, and for the pair (Egypt, 'ET'), where it is valid only for the time interval $[2009, 2013]$. Interestingly, for the pair (Egypt, 'ET'), the number of alone publications is higher than the number of internal and external ones in the time interval $[2003, 2010]$.

After this, we investigated the Herfindahl index. Obtained results generally confirm the corresponding aggregated ones (see Figure 12.6), although with some slight differences. In particular, the trend of 'NS' is identical. As for 'MH', differently from the aggregated case, Morocco always shows a behavior similar to the ones of Algeria and Tunisia. An intermediate behavior w.r.t. the ones of 'NS' and 'MH' is obtained for 'ET'.

The next investigation regarded backbones and cliques, clearly at the new disaggregation level. In this case, after having constructed the corresponding clique social networks, we found that the general trend (i.e., the aggregated results) about country backbones and hub cooperation in the different countries are confirmed in almost all research areas. However, there are a couple of interesting situations and/or exceptions. In fact, we can observe: *(i)* a very few number of hubs and cliques, along with a scarce cooperation, in Algeria for 'MH' and in Morocco for 'ET' in the time

interval [2003,2009], and in Morocco for 'MH' in the time interval [2007,2013]; *(ii)* the presence of two disconnected components in Algeria for 'MH' in the time interval [2007,2013] and in Morocco for 'MH' in the time interval [2003,2009]; *(iii)* a much more scarce cooperation for 'MH' than for 'NS' and 'ET' in all the countries and in both time intervals.

Analogously to the aggregated case, also at this disaggregation level we decided to construct the normalized *RA* social networks. To perform this task, we preliminarily had to verify that weight distribution in the edges followed a power law (see Section 12.4.3). After this, we constructed the normalized clique social networks and analyzed them. From this analysis we found that the general trends detected for aggregated data are still valid, although the following specificities/exceptions were observed: *(i)* there are few hubs and cliques and a scarce cooperation in Algeria for 'MH' in the time interval [2003,2009]; *(ii)* there is the presence of two disconnected components in Morocco for 'MH' in the time interval [2003,2009], and the presence of three disconnected components in Algeria for 'MH' in the time interval [2007,2013].

12.5.4 Investigation of the quality of publications

As for this issue, after having constructed G' (see Section 12.4.4), we computed the distributions of the metrics M_1, M_2 and M_3. We found that these distributions are analogous to the previous ones.

After this, we determined the number of hubs in the four countries of interest. In this case, we observed that, when considering the impact factor, the number of hubs generally decreases w.r.t. the case in which we consider only the number of publications (see Table 12.4). This is another indicator of the fact that the research performance for the four countries is low. In any case, we found that this number is always increasing over time.

Years	*Number of hubs* (without impact factors and citation numbers)	*Number of hubs* (with impact factors)	*Number of hubs* (with citation numbers)
2003	90	71	58
2004	85	70	65
2005	97	79	69
2006	100	95	82
2007	124	101	103
2008	140	118	111
2009	155	132	120
2010	165	142	123
2011	190	152	140
2012	199	157	127
2013	202	175	147

Table 12.4. Hub number over time in the three different situations into examination

As a final analysis, we computed the metrics M_1, M_2 and M_3 for each research area. We found that all the previous power law trends, obtained without considering impact factors, are fully confirmed and, even, reinforced, since the new power law distributions are steeper.

After the analyses based on impact factors, we made different analyses taking the citation number into account. In this case, we repeated all the computations already made for impact factors. In particular, we computed: *(i)* the distributions of the metrics M_1, M_2 and M_3; *(ii)* the number of hubs in the four countries (see Table 12.4); *(iii)* the distributions of the metrics M_1, M_2 and M_3 for each research area. All the results obtained in these computations totally confirm the ones seen for impact factors. Only the distributions of the metrics M_1, M_2 and M_3 present some noise near the elbow of the corresponding curves.

In our opinion, the fact that the previous results are confirmed in this case is extremely important, because this is an indicator of their stability; indeed, although we consider two totally different quality factors, the obtained results are always the same.

12.5.5 Characterization of hub neighborhoods

The first task we carried out for characterizing hub neighborhoods was the computation of the average number $AvgPub$ (see Section 12.4.5) of the publications of hub neighborhoods. As in Section 12.5.2, we distinguished among internal, external and alone publications. Obtained result is reported in Figure 12.11.

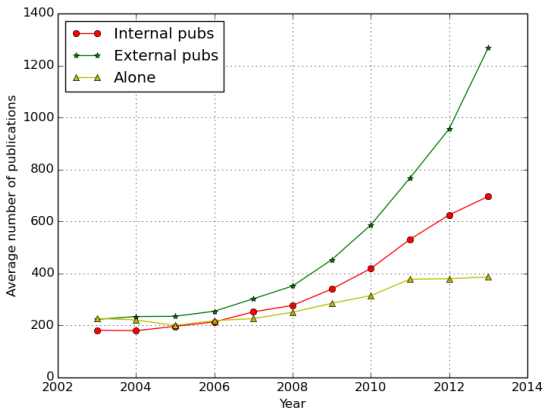


Fig. 12.11. Average number of internal, external and alone publications for hub neighborhoods

This result was quite unexpected. In fact, in Section 12.5.2, we have seen that hubs tend to publish more with foreign institutions than with internal ones. However,

in order to "fulfill its mission" to be a guide for its country, a hub must maintain a strict contact with internal institutions. So, we had hypothesized that this task was performed through its directed neighbors. Nevertheless, this graph seems to contradict this hypothesis.

As a matter of facts, a deeper investigation allows us to better understand this phenomenon. In fact, we must recall that, in the hub neighborhoods, there could be other hubs, which clearly can strongly influence the neighborhood behavior. As a consequence, it appears more correct to consider the trend of $\widehat{AvgPub}$, instead of $AvgPub$ (see Section 12.4.5), over time.

We carried out this last task by distinguishing among internal, external and alone publications. Obtained results, reported in Figure 12.12, fully confirm our hypothesis. In fact, in this case, the average number of internal publications is higher than the average number of external ones. Interestingly, the average number of alone publications is significant, at least from 2003 to 2010.

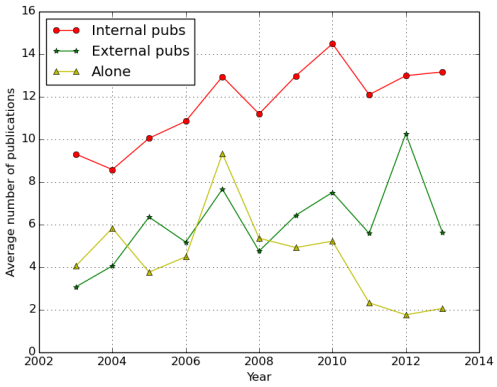


Fig. 12.12. Average number of internal, external and alone publications for hub neighborhoods (after the hubs present therein have been filtered out)

To verify if the four countries showed identical or different behaviors in this analysis, we disaggregated data per country and considered $AvgPub_k$ and $\widehat{AvgPub}_k$. We obtained that the trends described above for aggregated data are always confirmed for each country. We also observed an enormous decrease of the average number of publications when we consider the neighborhoods without hubs. This is a further confirmation that the distribution of the publications among institutions follows a power law. Finally, as usual, the number of alone publications performed by Egyptian hubs is quite significant.

A second investigation about neighborhoods regarded their average dimension. For this purpose, we computed $AvgDim$ over time. Obtained results are reported in Figure 12.13. From the analysis of this figure, we can observe that the average dimension of

hub neighborhoods always increases. This implies that the number of institutions cooperating with hubs is increasing over time constantly. As usual, we disaggregated these data per country. Specifically, we computed $AvgDim_k$ over time for the four countries. Obtained results evidence that the average dimension of neighborhoods increases in all countries, although with some irregularities.

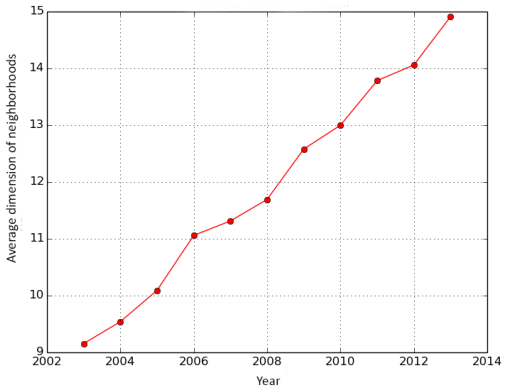


Fig. 12.13. Values of $AvgDim$ over time

After this, we investigated the cooperation level among the institutions belonging to hub neighborhoods. We started by computing $AvgCFrac$ over time. Obtained results are reported in Figure 12.14. Their analysis shows that $AvgCFrac$ tends to increase over time, although with some irregularities. This implies an increase over time of the cooperation among institutions belonging to hub neighborhoods. We also disaggregated data per country. For this purpose, we computed $AvgCFrac_k$ for the four countries. Obtained results evidence that the four countries show very different behaviors.

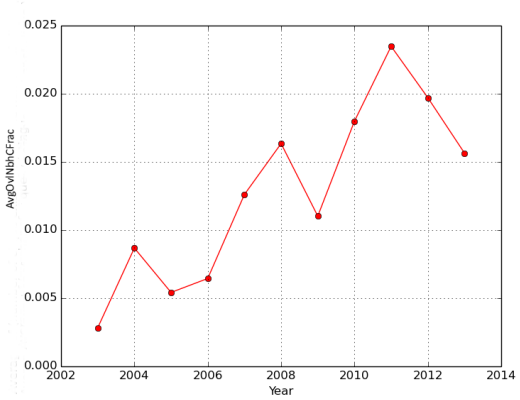


Fig. 12.14. Values of $AvgCFrac$ over time

A second measure about intra-neighborhood cooperation is $AvgCNbh$. We computed it over time. Obtained results are reported in Figure 12.15. From the analysis of this figure, we can observe that this parameter is significantly increasing over time, which is a further confirmation that cooperation among hub neighbors is increasing. In fact, its increase implies an increase of $NbhCNum_i$ and, since we have seen that $|nbh_i|$ increases over time, this implies a higher increase of the number of cliques. We disaggregated data per country. For this purpose, we computed $AvgCNbh_k$ for the four countries. Obtained results show that, in this case, the four countries present very different behaviors.

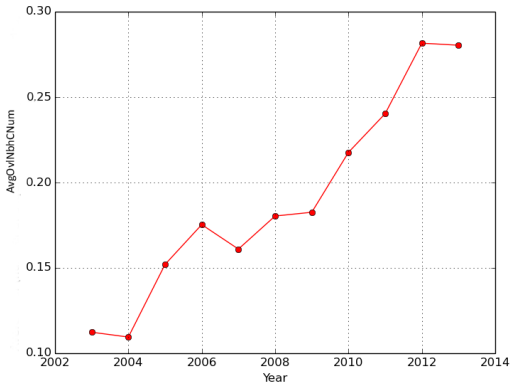


Fig. 12.15. Values of $AvgCNbh$ over time

The final measure about intra-neighborhood cooperation regarded the average density $AvgDens$ of the nbh social networks. We computed this parameter over time and we reported obtained results in Figure 12.16. From the analysis of this figure, we can observe an increase of this parameter. This is a further confirmation, obtained via a different fashion (based on edge number, instead of on clique number), that hub neighbors tend to increase their cooperation over time. We disaggregated data per country and we computed $AvgDens_k$ over time for the four countries. Obtained results show that this parameter is significantly increasing over time for all countries.

12.6 Discussion

In the Introduction, we have outlined the main innovations of our approach. Furthermore, in Section 11.2, we have examined related literature. Now, after having examined our approach in all details, and after having seen its behavior on a real case study, we can provide a more detailed presentation of its main features and novelties w.r.t. the previous ones.

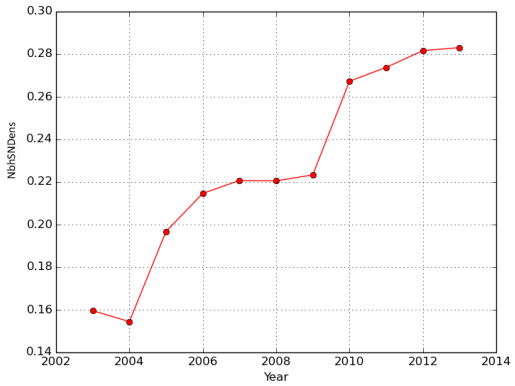


Fig. 12.16. Values of *AvgDens* over time

First, differently from most of the previous approaches described in Section 11.2, which focus on authors, our approach is centered on institutions.

One of the specific goals of our approach, i.e. hub detection and characterization, is novel in the literature. As matter of fact, to the best of our knowledge, the only paper investigating hubs is [36]. However, in [36], the definition of hubs is centered on authors and centrality measures, and is much simpler than the one we adopted in this chapter.

Our approach also aims at investigating the similarities and the differences of the research scenarios in a set of countries of interest. This is another contribution provided by it, which is generally not found in the previous approaches proposed in the past.

Also the techniques employed to carry out investigations are very different from the ones adopted in the past. In fact, past researches in this field were centered on the concept of centrality, whereas our approach employs more specific and ad-hoc data structures and parameters.

Furthermore, to better evaluate cooperation among involved institutions, we have employed the concept of clique and we have defined the clique social network, i.e., a specific support social network in which the dimension of a node is directly proportional to its tendence of cooperating with the other ones. We have also introduced some metrics to quantitatively evaluate the difference between two or more clique social networks.

Moreover, we have carried out a deep study of hub neighbors by introducing several metrics for quantitatively analyzing and comparing them.

As a further specificity of our approach, we have deepened our investigation about its main features, as well as about the similarities and the differences of research scenarios, by disaggregating data not only per country but also per research area and per pairs (country, research area).

In our evaluations, we considered not only the number of publications but also the corresponding quality by taking both their impact factor and their citation number into account.

Last, but not the least, we provided a re-definition of the Herfindahl index (largely used in the past research in Biology and Economics) to measure how much, in a given country, research activities are guided by few hubs or distributed among many institutions.

Another interesting issue could regard the comparison of our approach with some commercial systems, like Elsevier Pure, Elsevier Fingerprint Engine and Elsevier Scopus.

Elsevier Pure supports an institution in the definition of the optimal research and cooperation strategies, in assessment activities and in making business decisions. Pure aggregates information regarding the research activities of a given institution stored in different, both internal and external, sources. Furthermore, it ensures that data guiding strategic decisions is trusted, comprehensive and accessible in real time. It has an underlying centralized system, which is very versatile and supports the construction of reports, the evaluation of performances, the management of researchers' profile, the construction and the maintenance of research networks, the expertise detection, etc. Pure can be integrated with Elsevier Fingerprint Engine for stimulating the cooperation among researchers.

Elsevier Fingerprint Engine mines scientific documents ranging paper abstracts, funding announcements and awards, project summaries, patents, proposals/applications, etc., to create an index of weighted terms called Fingerprint visualization. The construction of Fingerprints is made through Natural Language Processing techniques and through the support of suitable thesauri. By aggregating and comparing the Fingerprints of people, publications, funding opportunities and ideas, Elsevier Fingerprint Engine mines metadata to detect connections among people, publications, funding opportunities and ideas. The thesauri adopted by Elsevier Fingerprint Engine make this last tool well suited in life science, engineering, earth and environmental sciences, arts and humanities, social sciences, mathematics and agriculture. Elsevier Fingerprint Engine can be integrated with Pure, to create expertise profiles aiming at helping cooperation, with Expert Lookup, to identify referees and potential conflicts of interest, and with Elsevier Journal Finder, to find the journal most suited to publish a given article.

Elsevier Scopus is the greatest database of abstracts and citations of scientific literature. It encompasses scientific journals, books and conference proceedings. Scopus supplies several functionalities. In particular, Scopus supports the search of documents, authors, affiliations and several forms of advanced search. It also allows the def-

inition of alerts regarding search, documents and authors, the browsing of resources, the creation of personalized lists of documents, the export of data to reference managers, the discovery of the documents citing selected articles, the visualization of the list of references included in an article, the analysis of search results, the comparison of journals, the quick visualization of the citation impact and the scholarly community engagement for an article, the analysis of the citation trend for an article, the analysis and tracking of an individual's citation history including total citations and document count, the computation of the h-index of an individual. Finally, Scopus has a comprehensive suite of metrics to facilitate evaluations and provide a better view of research interests.

Differently from Scopus, which bases most of its features on article citations, our approach is based on co-authorships. Furthermore, Scopus is more focused on single authors or single institutions, whereas our approach focus mainly on cooperations among authors or institutions. In its main objectives, our approach is more similar to Pure than to Scopus. However, Pure finds cooperation and network information based on text analysis performed by Fingerprint Engine. By contrast, as previously pointed out, our approach is based on co-authorship information.

Furthermore, our approach introduces the concept of hub, which is fundamental to help innovation managers in their decision making activities. This concept is not directly present in Pure and Fingerprint Engine. Only after several computations of the information directly provided by these systems, followed by a strong human intervention, it could be possible to derive (at least partial) information on hubs.

Analogously to Scopus, also our approach introduces several metrics to evaluate the level of the research activities in a scenario of interest, although the metrics used by the two systems take different pieces of information into account.

Once hubs have been detected, our approach allows a deep analysis of their main features and their relationships. For instance, it can indicate if there is a strong cooperation among the hubs of a given country or among the hubs (possibly of different countries) that operate in a given research area. Furthermore, it can investigate the characteristics of the hub neighbors and how they can be influenced by the hub themselves for the different countries and research areas of interest. Interestingly, our approach can incorporate in its metrics also citation counts and impact factors (i.e., the main parameters used by Scopus) to obtain more refined results.

13

Deriving knowledge on research scenarios in a set of countries

13.1 Deriving Knowledge

Patents have been one of the main topics investigated in several fields of scientific literature [8, 285, 446, 158, 141, 424]. In fact, they provide a wealth of useful information on the state of art and on the protagonists of a Research & Development (R&D) sector [466, 48, 156, 181, 200, 208, 272, 251, 308, 410, 285]. Patent submission is usually the first public claim of a new invention or innovation.

The investigation about both inventors and the patents submitted by them has appealed many researchers and economists, mainly in the last 15 years, and the interest on this topic has substantially increased over time.

Patent analysis can represent a useful tool for investigating the scientific development of a country. Moreover, understanding innovation evolution can allow decision makers to decide where it is better to concentrate investments. Furthermore, knowledge about patents allows decision makers to know the experiences of other (possibly competitor) organization/institutions/countries to verify the past and the current R&D activities and evolutions and to foresee the future ones. Finally, it provides a precise and detailed picture of the R&D cooperations between different organizations and/or countries and can represent an indicator of geo-political evolutions.

Several approaches for patent analysis have been proposed in the past. Most of them were based on classical statistics. However, currently, data about patents is rapidly increasing. As a consequence, the adoption of Data Mining and Big-Data-centered techniques appears compulsory. Among these last techniques, Social Network Analysis (SNA) appears particularly adequate [458, 53, 24, 106, 107, 258, 328, 476]. In fact, SNA allows the investigation of phenomena where involved data are huge and adopted variables are strictly related to each other, in which case classical statistical approaches present several difficulties to operate [439].

As a confirmation of the suitability of SNA for patent investigation, in the past, several approaches have been proposed in this sense (see, for instance, [48, 446, 158, 141, 208, 424]).

13.2 Approach description and knowledge pattern extraction

In this section, we present our approach, along with its support data structures and parameters, and we show how it can answer the ten research questions mentioned in the Section 13.1. In this chapter we used same data presented in Section 12.3 However, before starting our presentation, we must define some sets allowing the formalization of data at our disposal.

The first set regards IPC classes. It consists of the following elements:

$$IPC = \{\text{“ICT”, “INS”, “CM”, “PB”, “IP”, “ME”, “CE”}\}$$

where “ICT” (risp., “INS”, “CM”, “PB”, “IP”, “ME”, “CE”) denotes “Information and Communication Technologies” (risp., “Instruments”, “Chemicals and Materials”, “Pharmaceuticals and Biotechnology”, “Industrial Processes”, “Mechanical Engineering”, “Civil Engineering”).

The second set, called Pat, represents all the patents registered in our database. Given a patent $p \in Pat$, we indicate with $Inventors_p$ the set of its inventors, and with $Classes_p$ the set of the IPC classes it belongs to.

Now, we can define:

- the set Pat_k of the patents filed by at least one inventor of the country k;
- the set Pat^q of the patents belonging to the IPC class q;
- the subset Pat_k^q of Pat_k whose elements refer to the IPC class q.

Finally, we define a social network that represents our main support data structure for our investigations. Specifically, it is represented by:

$$G = \langle N, E \rangle$$

N indicates the set of nodes of G. A node $n_i \in N$ corresponds to exactly one inventor registered in our database. Since there is a biunique correspondence between a node of N and the corresponding inventor, in the following we will use the symbol n_i to denote both of them. A label is associated with each node of N; it represents the country of the corresponding inventor. We denote by l_i the label of n_i. E is the set of the edges of G. There exists an edge $e_{ij} = (n_i, n_j, w_{ij}) \in E$ if there exists at least one patent filed by both n_i and n_j. w_{ij} is the weight of e_{ij}; it denotes the number of patents filed by both n_i and n_j.

Starting from this support data structure, we can define some sets representing the neighborhood of a node in G. Specifically, we define the direct neighborhood nbh_i of a node $n_i \in G$ as the set of the nodes of G directly connected to n_i. Then, we can define the set nbh_i^I (resp., nbh_i^F) of the direct neighbors of n_i belonging to the same country as (resp., a country different from) the one of n_i. Finally, we define the set N_k of the nodes (i.e., inventors) of a country k.

As previously pointed out, the amount of data to process, along with the objective to define a general approach that can be adopted also in the future (when available data about patents will enormously grow), led us to exploit Big Data technology. In particular, we adopted the MongoDB [6] DBMS. We also adopted NetworkX [1], a powerful Python library providing all the basic algorithms for SNA. NetworkX can interact with MongoDB via Python. Thanks to the flexibility and the power of this last language, we could easily exploit the basic SNA functions provided by NetworkX to construct the (often very complex) algorithms underlying our parameters and knowledge extraction tasks.

In the following, we answer the ten research questions, one per subsection. In carrying out this task, we introduce new metrics, parameters and data structures; we also derive several knowledge patterns.

13.2.1 RQ1: What is the distribution of patents against inventors

To answer this question, given a node n_i, we defined a metric M_1 such that M_{1_i} denotes how much patents were filed by the inventor n_i. This metric coincides with the classical weighted degree centrality [193]. We measured M_1 for the following countries: all European Union countries, all Mediterranean countries, all North African countries, all countries of BRICS (Brazil, Russia, India, China and South Africa), South Korea, Japan, Vietnam and Taiwan.

For most of these countries, M_1 follows a power law distribution. This implies that, in the corresponding countries, there are few inventors filing many patents and many inventors filing very few patents. For instance, in Figure 13.1, we show the distribution of M_1 for France.

Some countries (i.e., Greece, Croatia, Principate of Monaco, Slovenia, Turkey and all countries of BRICS) show a slightly disturbed power law. For instance, in Figure 13.1, we show the distribution of M_1 for Greece. This result is motivated by the fact that, in these countries, the maximum number of inventors and filed patents is quite low.

Other countries (i.e., Egypt, Lebanon, Malta, Morocco and Tunisia) present a disturbed power law. For instance, in Figure 13.2, we show the distribution of M_1 for Egypt. Also in this case, the obtained trend is justified by the low number of inventors

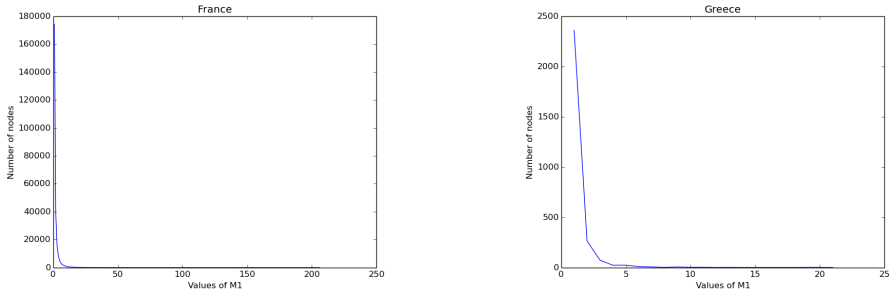


Fig. 13.1. Distribution of M_1 for France and Greece

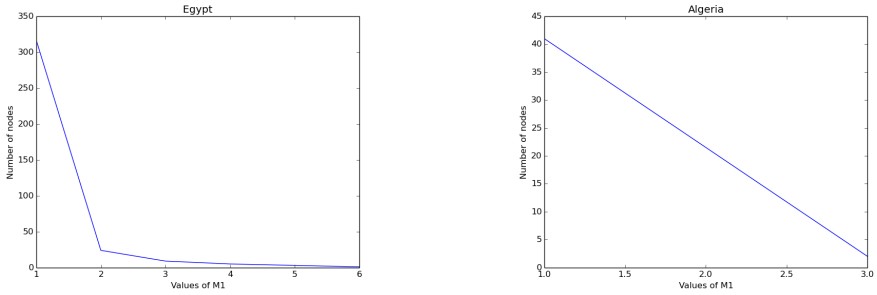


Fig. 13.2. Distribution of M_1 for Egypt and Algeria

and filed patents in the corresponding countries. Finally, for some countries (i.e., Albania, Algeria, Libya and Montenegro), the distribution of M_1 is totally different from a power law. In some cases, even a linear distribution can be observed. For instance, in Figure 13.2, we show the distribution of M_1 for Algeria. For all the cases in which the distribution of M_1 is totally different from a more or less disturbed power law, the number of inventors and filed patents is so scarce to make little significant and unreliable any investigation about them.

13.2.2 RQ2: How the number of inventors and their cooperation degree evolve over time?

To answer this question, given a country k, first we constructed a support social network:

$$G_k = \langle N_k, E_k \rangle$$

The set N_k was defined previously. There exists an edge $e_{ij} = (n_i, n_j) \in E_k$ if both n_i and n_j belong to N_k.

After this, we computed the temporal evolution of the number of nodes $|N_k|$, the number of edges $|E_k|$ and the density D_k for G_k.

The computation of these parameters is important to understand the temporal evolution of both the inventors of a given country and their collaborations. Generally,

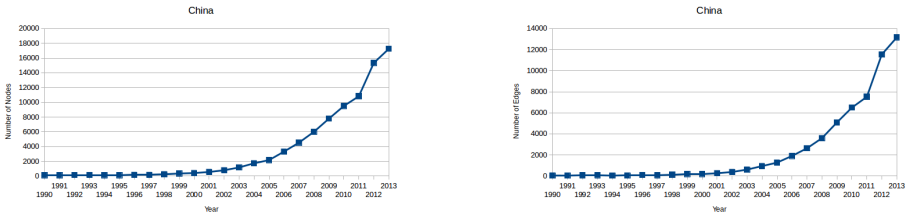


Fig. 13.3. Trend of $|N_k|$ and $|E_k|$ over time for China

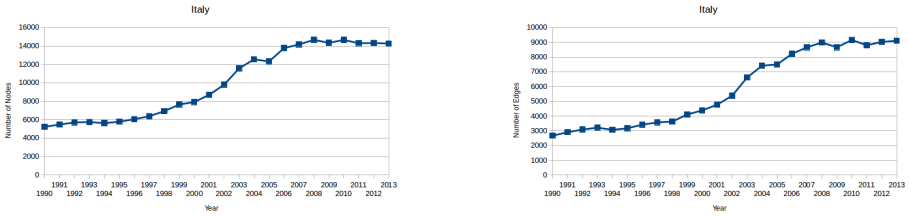


Fig. 13.4. Trend of $|N_k|$ and $|E_k|$ over time for Italy

all countries into consideration present a growing number of nodes and edges against years. Some countries (e.g., those of BRICS, South Korea, Taiwan, Turkey and Israel) present an exponential growth. For instance, in Figure 13.3, we show the trend of the number of nodes for China. From the analysis of this figure, we can observe that this number grows exponentially, with a sudden rise from 2002. If we continue to analyze China, and we consider the number of edges against time (quantifying the level of cooperation among inventors), we observe an analogous trend, i.e., an exponential rise starting from 2002, as shown in Figure 13.3.

Most of the European Union countries, instead, show an increasing linear trend. For instance, in Figure 13.4, we show the trend of the number of nodes for Italy. From the analysis of this figure, we can observe that this significantly grows up to 2008. After this year, the trend is roughly constant. In an analogous fashion, also the number of edges shows an increasing linear trend up to 2008 and has stalled from 2008 onwards (see, again, Figure 13.4). North African countries and some European Union ones show, instead, a growing but irregular trend.

As for Density D, it generally decreases. This behavior can be explained by the fact that the value of the density of a network is inversely proportional to the square of the number of nodes. To obtain a constant trend against time, it would be necessary that the number of edges grows proportionally to the square of the number of nodes, which is unthinkable in real scenarios. As an example, in Figure 13.5, we show the trend of D for China and Italy. Observe that the decrease is more marked for China than for Italy. This is justified by the fact that the increase of the number of nodes was exponential for China and linear for Italy (see, Figures 13.3 and 13.4).

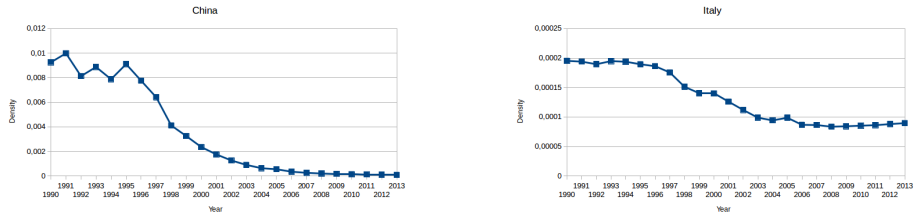


Fig. 13.5. Trend of $|D_k|$ over time for China and Italy

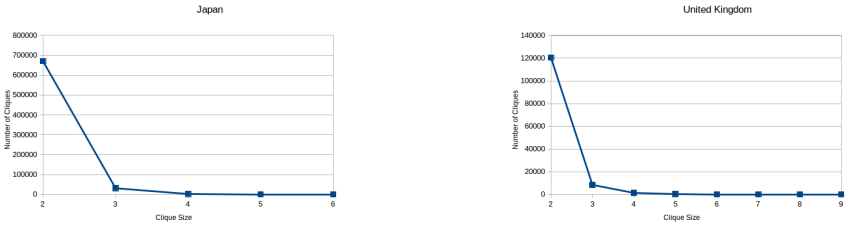


Fig. 13.6. Distribution of clique size for Japan and UK

13.2.3 RQ3: Do cliques of inventors exist in some countries?

In Social Network Analysis, a clique is a subgraph where every node is adjacent to every other. In our scenario, it is an indicator of a compact group of inventors, who cooperate each other intensively.

To answer RQ3, we initially computed the distribution of the dimension of cliques of the social network G_k for the countries of interest. In fact, since, in a clique, all nodes are totally connected to each other, the dimension of a clique can be considered as a valid metric to understand how much the inventors of a given country tend to form more or less large work groups. The general trend we found for this phenomenon is the one of a power law with a dimension of the maximum clique different in the different countries. For instance, in Figure 13.6, we report the distribution of cliques for Japan and United Kingdom. In both cases, there is a power law distribution, with a different maximum number of cliques. In fact, in spite Japan has a number of nodes much higher than UK, the dimension of its maximum clique is lower than the one of UK.

A case of particular interest is represented by Israel, which presents a maximum dimension of cliques equal to 7. This is a very high value if we consider that the number of Israelis inventors (i.e., 30,358) is much lower than the one of Japan (i.e., 924,554) and UK (i.e., 231,128) inventors.

To capture and quantify this last observation, we decided to define a parameter that indicates how much the inventors of a country k are aggregated in cliques. The previous result suggests that the difficulty to have larger and larger cliques grows

Country	$\lvert N_k \rvert$	$\lvert \mathcal{C}_k \rvert$ $(\nu_{\lvert \mathcal{C}_k \rvert})$	$\lvert \mathcal{C}_k \rvert - 1$ $(\nu_{\lvert \mathcal{C}_k \rvert - 1})$	$\lvert \mathcal{C}_k \rvert - 2$ $(\nu_{\lvert \mathcal{C}_k \rvert - 2})$	Agg_k
Brazil	7721	5(1)	4(8)	3(118)	0.143
Russia	12813	7(2)	6(5)	5(16)	0.085
Taiwan	18729	7(7)	6(20)	5(118)	0.317
India	26516	6(1)	5(0)	4(101)	0.063
Israel	30358	7(8)	6(12)	5(42)	0.103
Denmark	30762	9(1)	8(0)	7(22)	0.108
Finland	31903	7(1)	6(17)	5(72)	0.110
Austria	40734	7(7)	6(20)	5(118)	0.146
Spain	43131	9(6)	8(13)	7(41)	0.270
Belgium	48073	8(1)	7(4)	6(33)	0.059
China	54419	8(5)	7(17)	6(37)	0.107
Sweden	63593	9(10)	8(5)	7(6)	0.113
South Korea	115272	9(2)	8(16)	7(58)	0.109
United Kingdom	231128	9(4)	8(11)	7(7)	0.025
Japan	924554	6(26)	5(244)	4(2294)	0.049

Table 13.1. Values of Agg_k for several countries

exponentially. Therefore, to define a corresponding index, we judged suitable to consider only the cliques of maximum, sub-maximum and sub-sub-maximum dimension, as well as to assign an exponentially decreasing weight to these cliques. Specifically, to define the aggregation index Agg_k of inventors on cliques for the country k, we have preliminarily defined: *(i)* the maximum clique $\mathcal{C}_k$ of the country k; *(ii)* the dimension $|\mathcal{C}_k|$ of this clique;*(iii)* the number ν_x of cliques having dimension x. Agg_k is defined as:

$$Agg_k = \frac{\sum_{l=0}^{2} 2^{(|\mathcal{C}_k|-l)} \cdot \nu_{(|\mathcal{C}_k|-l)}}{|N_k|}$$

In Table 13.1, we report the dimension of the maximum, sub-maximum and sub-sub-maximum cliques for several countries, along with the value of the corresponding aggregation index. We decided to organize this table in such a way as to group countries having a similar number of inventors. For this reason, we divided it in three parts: the first for countries having less than 35,000 inventors, the second for countries having a number of inventors between 40,000 and 65,000, and the third for countries having more than 110,000 inventors.

In the first sub-table, the country having the highest aggregation index is Taiwan. This result can be explained by observing that, even if Taiwan has a number of nodes quite low w.r.t. that of the other countries of this sub-table, it has a high number of cliques of maximum, sub-maximum and sub-sub-maximum dimension. In this sub-table, India has the lowest value of Agg_k; in fact, despite the number of cliques of sub-sub-maximum dimension is high, there is only one clique of maximum dimension and no clique of sub-sub-maximum dimension. Analogous reasonings can be made for the second sub-table, where the country with the highest (resp., lowest) value of Agg_k is Spain (resp., Belgium). As for the third sub-table, we can observe that the values of Agg_k are generally small. The highest value is reached by South Korea that, even

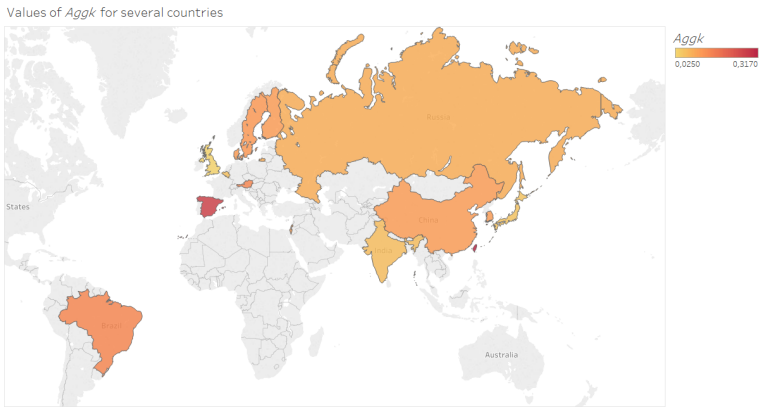


Fig. 13.7. Visualization of the values of Agg_k for the countries reported in Table 13.1

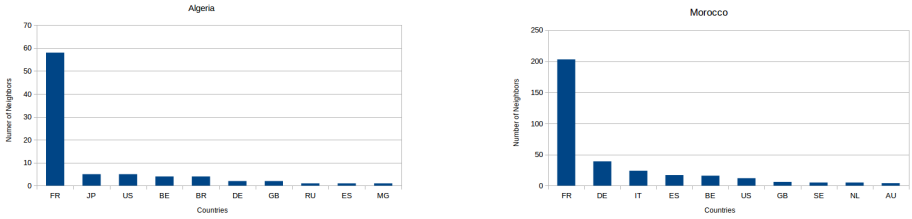


Fig. 13.8. Distribution of foreign collaborations for Algerian and Moroccan inventors

if it has much less inventors than UK and Japan, presents a dimension of maximum clique equal to the one of UK and even higher than the one of Japan. In Figure 13.7, we report a map providing an intuitive visualization of the values of Agg_k for the countries reported in Table 13.1.

13.2.4 RQ4: With whom and how inventors cooperate?

This RQ aims at analyzing the distribution of the countries with which a given country mostly cooperates (at least, as far as patents are concerned). To answer this question, given a country k, for each node $n_i \in N_k$, we considered nbh_i^F. Then, we computed the distribution of countries which the nodes associated with these neighborhoods belonged to. We performed this analysis for all Mediterranean countries, for the countries of BRICS and for some European Union and North African countries. In particular, we focused our attention mainly on some past colonies. In Figures 13.8 and 13.9, we show the results obtained for some French past colonies (e.g., Algeria, Morocco and Tunisia) and for a British past colony (e.g., Egypt).

As one would expect, most of the inventors of Algeria, Morocco and Tunisia cooperate mainly with French inventors. As a further interesting observation, for both Morocco and Tunisia, beside cooperation with France, there is a significant cooperation with Germany. This does not happen for Algeria, where the distribution with non-French inventors is "pulverized". Now, we focus on Egypt, which (we recall)

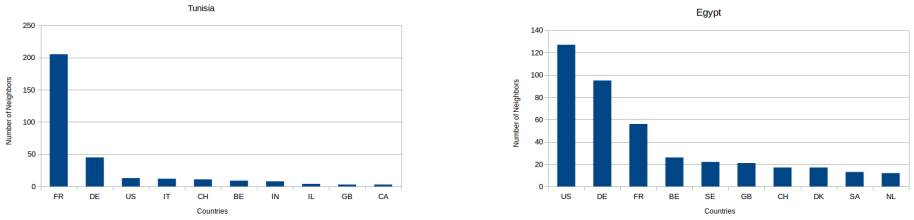


Fig. 13.9. Distribution of foreign collaborations for Tunisian and Egyptian inventors

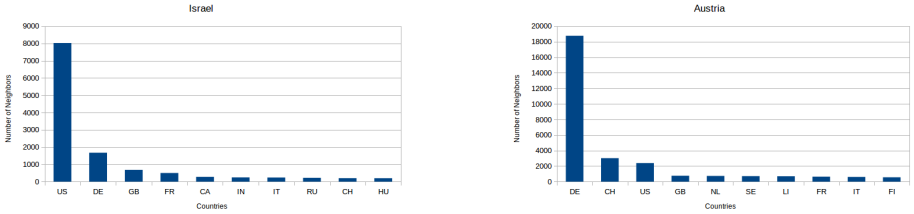


Fig. 13.10. Distribution of foreign collaborations for Israelis and Austrian inventors

is a British past colony. Differently from Algeria, Morocco and Tunisia, we observe that Egyptian inventors do not show a prevalent collaboration with inventors of their motherland (i.e., United Kingdom - GB in Figure 13.9). In fact, the highest number of collaborations can be found with inventors of United States, Germany and France. Interestingly, there is a good contribution with inventors coming from Saudi Arabia. This result is a confirmation of a study about research cooperation in North African countries reported in [251].

In Figure 13.10, we show the results obtained for Israel. In this case, we can observe a strong cooperation with US inventors and, beside them, with German ones.

In Figures 13.10 and 13.11, we show the results obtained for Austria and Slovenia. From the analysis of this figures it is possible to observe that, for both these countries, most of foreign collaborations are performed with German inventors. In particular, as for Slovenian inventors, we can observe a high concentration of collaborations with inventors coming from Germany and Austria (as evidence of the very strong links between Slovenia and German-speaking countries dating to the period of the Austro-Hungarian Empire) and, to a lesser extent, with inventors from Croatia.

Finally, we focused on Taiwan. The corresponding results are shown in Figure 13.11. From their analysis, we can see that the country with which Taiwan inventors mostly cooperate is China. This result is quite surprising if we consider the quite conflicting political relations between these two countries after the Second World War.

After this first activity, we computed the variety level of the countries, which the inventors of a country k cooperate with. Drawing on a measure of biodiversity

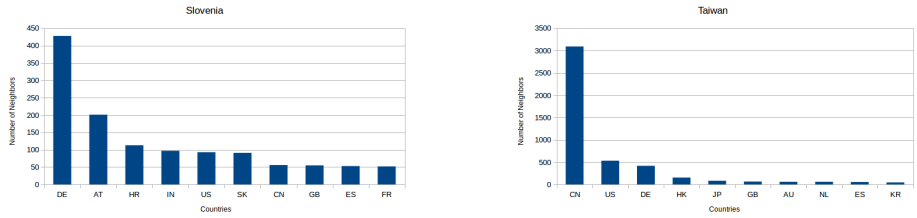


Fig. 13.11. Distribution of foreign collaborations for Slovenian and Taiwan inventors

introduced by Simpson (1949), we build an indicator of the internationalization level of the inventor teams relying on cross-patent data. Such indicator was used to conduct an explorative empirical analysis of the trends and the features in research groups' internationalization level using data from the Worldwide Patent Statistical Database. It was also adopted to measure the size of firms in relation to the industry and the amount of competition among them, and is known as Herfindahl Index. As far as this research question is concerned, the Herfindahl Index HI_k indicates if the inventors of a given country privilege collaborations with the inventors of one or more foreign countries. The higher the Herfindahl Index of a country k, the more concentrated the external collaborations of k.

The values of the Herfindahl Index obtained for several countries are reported in the second column of Table 13.2.

Among the North African countries, the highest value of HI is obtained by Algeria. This result is due to the fact that Algerian inventors cooperated mainly with the inventors of one country, i.e., France. An analogous observation can be drawn for Tunisia and Morocco. Differently from these three countries, Egypt has a much lower value of HI, because the collaborations of Egyptian inventors are more distributed among several countries. Analogous reasonings can be drawn for European Union countries (where the highest value of HI is obtained by Austria). As for the countries of BRICS, we can observe that the values of HI are generally small. In fact, Brazil, South Africa, China and India tend to cooperate with several countries. High values of HI can be obtained also for Taiwan, Israel and Turkey (this last country cooperates mainly with USA). In Figure 13.12, we report a map providing an intuitive visualization of the values of HI for the countries reported in Table 13.2.

Actually, this definition of the Herfindahl Index has a weak point in that the obtained value could be strongly distorted by the presence of a large number of extemporaneous collaborations between an inventor of k and an inventor of another country, who cooperated for one or two patents only. Taking the power law trend, typical of the measures in our reference scenario, we thought to use a modified version of the Herfindahl Index obtained by limiting the countries into consideration to

Country	HI_k	HI_k *Top 80%*	HI_k^* *Top 80%*
Egypt	0.144	0.148	0.147
Morocco	0.358	0.367	0.365
Tunisia	0.434	0.442	0.441
Algeria	0.502	0.527	0.521
Libya	0.180	0.250	0.143
Spain	0.138	0.138	0.137
Estonia	0.128	0.130	0.129
Poland	0.118	0.119	0.119
Austria	0.343	0.343	0.343
Bulgaria	0.176	0.179	0.179
Romania	0.173	0.175	0.175
Slovenia	0.095	0.096	0.095
Croatia	0.132	0.135	0.134
Brazil	0.138	0.138	0.137
India	0.128	0.130	0.129
China	0.118	0.119	0.119
South Africa	0.159	0.160	0.160
Israel	0.369	0.370	0.370
Taiwan	0.415	0.442	0.441
Turkey	0.253	0.255	0.254
Bosnia-Herzegovina	0.176	0.187	0.156

Table 13.2. Values of HI, HI Top 80% and HI^* Top 80% for North African countries

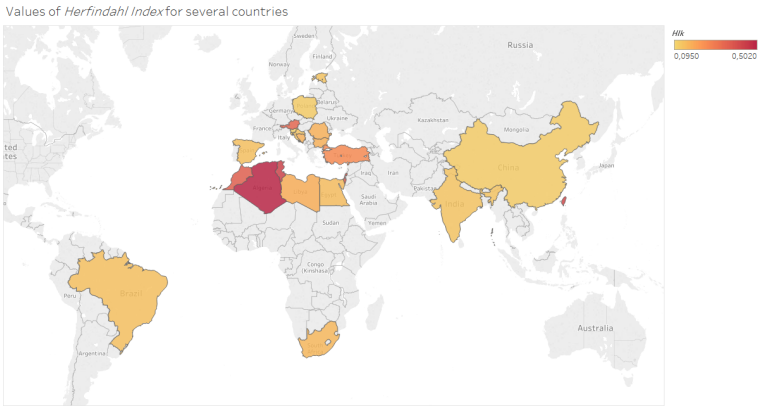


Fig. 13.12. Visualization of the values of HI for the countries reported in Table 13.2

the top 80%. In this way, the tail of the power law distribution, which is the main cause of the distortions mentioned above, is removed. The obtained results are reported in the third column of Table 13.2.

From the analysis of this table, we can observe that values obtained in this way are slightly higher than the ones obtained previously. Libya represents an exception in these results since, for this country, the values obtained with this new definition are much higher than the ones obtained previously. This fact is due to the extremely low number of Lybian patents, which makes any result about this country unreliable.

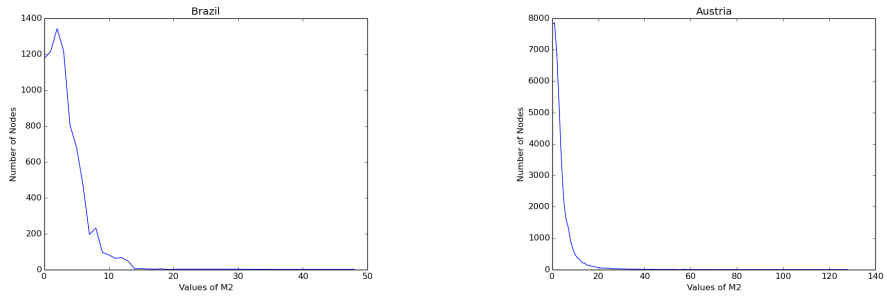


Fig. 13.13. Distribution of M_2 for Brazil and Austria

Nevertheless, a problem is still present. In fact, owing to its definition, the Herfindahl Index ranges in the interval $\left[\frac{1}{0.8 \cdot |FCntr_k|}, 1\right]$, where $FCntr_k$ denotes the set of the countries having at least one inventor that filed a patent with at least one inventor of the country k. As a consequence, the range of values of this index differs from one country to another. Owing to this fact, the comparison of the values of the Herfindahl Index for different countries could produce distorted results. To avoid this problem, and to make sure that the range of the values of the Herfindahl Index is identical for all countries, we decided to adopt the modified version of the Herfindahl Index HI_k^* proposed in [190]. The obtained results are reported in the four column of Table 13.2. For most countries, obtained values of HI^* are very similar to (more precisely, slightly lower than) the ones obtained after the first refinement. Also in this case, Libya is an exception. The reason for this fact is the same as the one we discussed previously.

13.2.5 RQ5: What about the "neighbors" of inventors?

To answer this question, first we defined a new metric M_2 such that M_{2_i} indicates the dimension of the neighborhood of n_i. M_2 is effective for helping us to understand how much the inventors of a given country k tend to cooperate for filing patents.

We measured this metric for the same countries considered in $RQ4$. For most of them, it follows quite a disturbed power law distribution. For instance, in Figure 13.13, we show the distribution of M_2 for Brazil. It presents a peak for a value of M_2 between 0 and 10. The only country presenting a perfect power law distribution is Austria, as shown in Figure 13.13.

After this, we investigated the level of cooperation with foreign colleagues for the inventors of a given country k. For this purpose, we defined a metric M_3 such that M_{3_k} denotes the average fraction of foreign collaborations carried out by investors of k. Actually, M_3 allows us to understand how much the inventors of k tend to cooperate with foreign colleagues. It ranges between 0 and 1; the higher M_3 the higher the tendency of the inventors of k to cooperate with foreign ones.

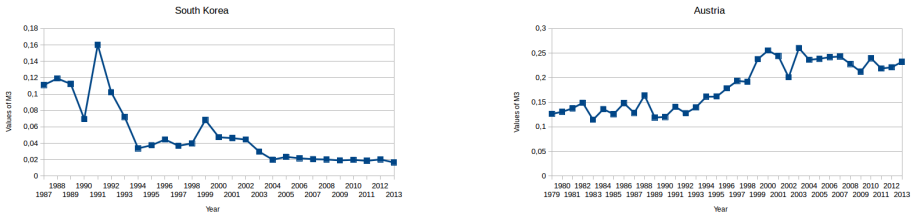


Fig. 13.14. Trend of M_3 over time for South Korea and Austria

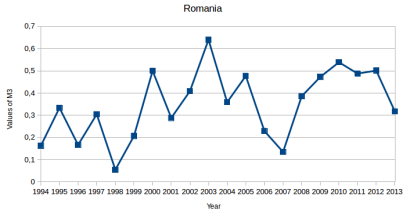


Fig. 13.15. Trend of M_3 over time for Romania

We computed the trend of M_3 over time for several countries to understand if (and, in the affirmative case, how much) a given country has become international over time. As for this analysis, we considered some European Union countries, some countries of BRICS, some North African countries, South Korea and Taiwan.

Obtained results are very heterogeneous. Some countries (e.g., Spain, Taiwan, South Korea, China and Brazil) present a generally decreasing trend. As an example, in Figure 13.14, we report the trend of M_3 for South Korea. From the analysis of this figure, we can observe that, in the last years, South Korean inventors tend to cooperate more and more with internal ones. Other countries (e.g., Austria, Italy and South Africa), instead, present an increasing trend for this measure. For instance, in Figure 13.14, we show the trend of M_3 for Austria. From the analysis of this figure, we can observe that, in the last 20 years, Austrian inventors tend to internationalize more and more. Finally, other countries present quite an irregular trend for M_3, characterized by the presence of peaks and decays over time. For instance, in Figure 13.15, we show the trend of M_3 for Romania.

13.2.6 RQ6: Do power inventors exist?

To answer this question, we had to preliminarily specify who is a "power inventor". With regard to this definition, we point out that we do not aim at proposing a new concept characterized by a mathematical foundation supporting it. Instead, we would like to introduce an informal and empirical, yet reasonable, concept, which can give a picture about this phenomenon and can support the extraction of innovation

geography knowledge patterns about it. Taking this premise into account, we defined a "power inventor" as an inventor that fulfills the following conditions:

- C_1: she files many patents;
- C_2: she has a lot of collaborations;
- C_3: she has an international feel, which implies that she cooperates very much with inventors of foreign countries.

The reasoning underlying these conditions are the following:

- If an inventor filed very few patents, even if all of them were in cooperation with foreign inventors, she cannot have such a weight to influence the scenario of her country.
- If an inventor filed many patents, but all of them were in cooperation with few inventors, she would not have the capability (fundamental for power inventors) to stimulate, through collaborations, other inventors to file patents.
- If an inventor filed many patents, but all of them were in cooperation with inventors of her countries, she would certainly be an important protagonist in her country, but she would not have the capability (fundamental for power inventors) to stimulate contacts with foreign countries.

To the best of our knowledge, the term "power inventor", or a concept analogous to it, has not been previously introduced in the literature.

A metric for evaluating condition C_1 (resp., C_2) is the metric M_1 (resp., M_2) defined in Section 13.2.1 (resp., Section 13.2.5). A metric for evaluating condition C_3 (we call M_4 this metric in the following) is analogous to M_2 except that nbh_i is substituted by nbh_i^F, since C_3 is focused on collaborations with foreign countries. Interestingly, M_2 and M_3 are analogous to E-I index [193].

We measured M_4 for the Euro-Mediterranean countries, the North African ones and, finally, the countries of BRICS. For most of these countries, M_4 follows a power law distribution. As an example, in Figure 13.16, we show the distribution of M_4 for France. Some countries (e.g., Greece, Turkey and Principate of Monaco) present a slightly disturbed power law distribution for M_4. For instance, in Figure 13.16, we show the distribution of M_4 for Greece. For other countries, such as Morocco, Slovenia, Tunisia, Croatia, Cyprus and Egypt, M_4 follows a disturbed power law distribution. For instance, in Figure 13.17, we show the distribution of M_4 for Croatia. Finally, for other countries, like Syria, Albania, Algeria, Bosnia-Herzegovina, Lebanon, Libya and Malta, M_4 does not follow a power law distribution. As an example, in Figure 13.17, we show the distribution of M_4 for Malta. This last case refers to countries having a very low number of collaborations with foreign countries, which makes any analysis for them not reliable.

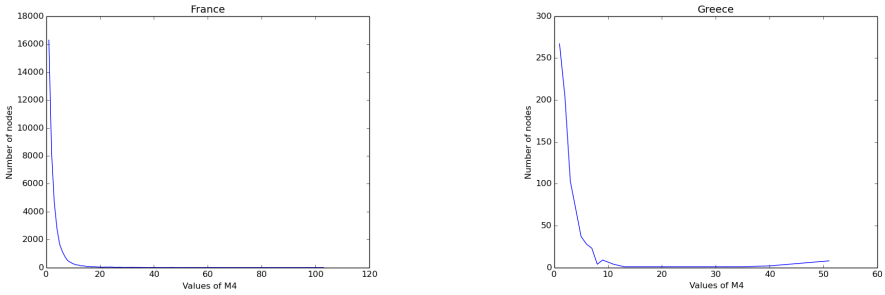


Fig. 13.16. Distribution of M_4 for France and Greece

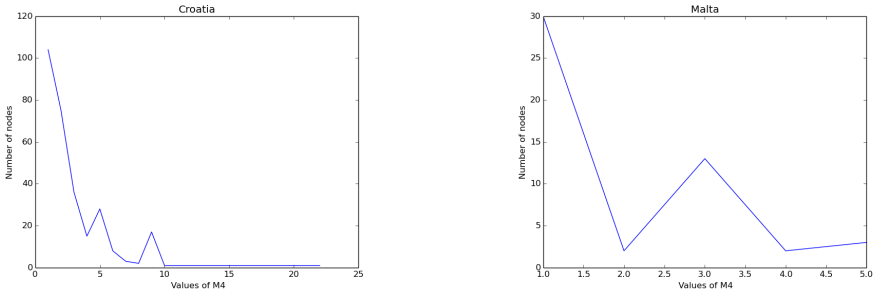


Fig. 13.17. Distribution of M_4 for Croatia and Malta

Taking all these considerations into account, the set $\mathcal{P}_k^X$ of the power inventors of a country k can be defined as the set of those inventors simultaneously belonging to the top $X\%$ of the inventors with the highest values of M_1, M_2 and M_4.

As previously pointed out, M_1, M_2 and M_4 generally follow a power law (possibly disturbed) distribution. Since available data are huge, and since the power law distributions characterizing M_1, M_2 and M_4 are generally steep, we decided to consider a low value for X and we set $X = 5$. As a consequence, in the following, when X is not specified, we intend that it is equal to 5.

13.2.7 RQ7: Does a backbone of power inventors exist?

To answer this question, first we constructed a support data structure that we called *power inventor social network*. Given a country k, it is defined as:

$$G_k^{\mathcal{P}} = \langle \mathcal{P}_k, E_k^{\mathcal{P}} \rangle$$

where $\mathcal{P}_k$ is the set of the power inventors of k (see Section 13.2.6) and $E_k^{\mathcal{P}} = \{(n_i, n_j, w_{ij}) | n_i, n_j \in \mathcal{P}_k,\ w_{ij}$ is the number of patents filed by n_i and n_j together$\}$. Then we computed:

- the aggregation index Agg_k related to the inventors of k;
- the aggregation index $Agg_k^{\mathcal{P}}$ related to the power inventors of k;

Country	$rAgg_k$	rf_k
Brazil	5.595	6.763
Taiwan	29.012	6.541
China	23.483	5.846
Austria	44.084	4.193
Italy	47.472	5.100
Israel	22.583	3.674
Russia	10.034	5.149
Spain	25.498	6.624

Table 13.3. Values of $rAgg_k$ and rf_k for some countries

- the fraction f_k of the nodes of N_k belonging to at least one clique of G_k having a dimension greater than or equal to 3;
- the fraction $f_k^{\mathcal{P}}$ of the nodes of $\mathcal{P}_k$ belonging to at least one clique of $G_k^{\mathcal{P}}$ having a dimension greater than or equal to 3;
- the ratio $rAgg_k = \frac{Agg_k^{\mathcal{P}}}{Agg_k}$;
- the ratio $rf_k = \frac{f_k^{\mathcal{P}}}{f_k}$.

The values of these last two parameters for some countries are reported in Table 13.3.

We recall that the aggregation index is an indicator of how much the inventors of a country k are aggregated in cliques. By examining Table 13.3 we can see that, for countries like Austria and Italy, which present the highest value of $rAgg_k$, the corresponding power inventors are much more aggregated in cliques than the inventors of their countries. As for rf_k, we can note that the corresponding values are quite homogeneous and range between 3 and 7. Since, for all countries, both $rAgg_k$ and rf_k present values higher or much higher than 1, we can undoubtedly conclude that, in each country, there exists a backbone among power inventors. In Figure 13.18, we report a map providing an intuitive visualization of the values of $rAgg_k$ for the countries reported in Table 13.3.

.

To deepen our research about this issue, we defined a new data structure called *clique social network*. In particular, let $\mathcal{CS}_k$ be the set of cliques of $G_k^{\mathcal{P}}$ having dimension greater than or equal to 3. A clique social network $CG_k^{\mathcal{P}}$, corresponding to $G_k^{\mathcal{P}}$, is defined as:

$$CG_k^{\mathcal{P}} = \langle \mathcal{P}_k, CE_k^{\mathcal{P}} \rangle$$

$CG_k^{\mathcal{P}}$ has a node for each power inventor of $\mathcal{P}_k$ that belongs to at least one clique of $\mathcal{CS}_k$. Each node n_i of $CG_k^{\mathcal{P}}$ has associated a weight denoting the number of cliques

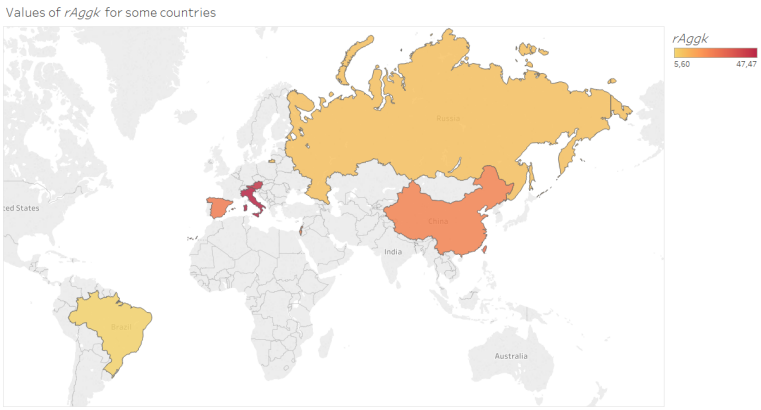


Fig. 13.18. Visualization of the values of $rAgg_k$ for the countries reported in Table 13.3

Fig. 13.19. The clique social network of Spain and a zoomed portion of it

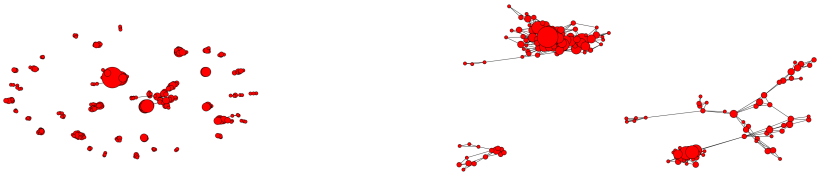


Fig. 13.20. The clique social network of Israel and a zoomed portion of it

of $\mathcal{CS}_k$ which it belongs to. An edge $e_{ij} = (n_i, n_j) \in CE_k^{\mathcal{P}}$ indicates that n_i and n_j simultaneously belong to at least one clique of $\mathcal{CS}_k$.

In Figure 13.19, we present the clique social network of Spain. As shown in this figure, we can observe that this social network is characterized by a particular dense core denoting that, in this country, there is a group of particularly active inventors, who often cooperate each other. In Figure 13.20, we show the clique social network of Israel. In this case there is not a dense core but, on the contrary, there are several nodes whose dimension is generally smaller than the ones belonging to the core of Spain.

Now, we want to define some parameters providing a quantitative measure of what an expert could visually capture by observing clique social networks. In particular, in Table 13.4, we report the number of nodes, the number of edges and the density of the clique social networks for some countries.

Country	*Number of Nodes*	*Number of Edges*	*Density*
Brazil	19	23	0.135
India	209	417	0.019
China	577	1299	0.008
Israel	336	744	0.013
France	3452	8615	0.001
Italy	2125	5763	0.003
Spain	618	1672	0.009

Table 13.4. Number of nodes, number of edges and density of the clique social networks of some countries

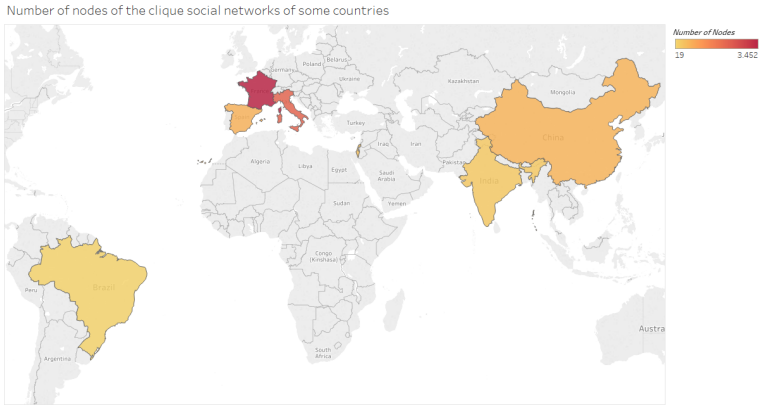


Fig. 13.21. Visualization of the number of nodes of the clique social networks of the countries reported in Table 13.4

From the analysis of this table, we can see, for instance, that France and Italy have the lowest values of density. This can be explained by observing that both these two countries have many nodes only partially connected to each other. By contrast, Brazil has a very low number of nodes and edges but, at the same time, the highest density among the countries into consideration. This denotes that Brazilian power inventors are strongly connected to each other. In Figures 13.21, 13.22 and 13.23, we report some maps providing an intuitive visualization of the number of nodes, the number of edges and the density of the clique social networks of the countries reported in Table 13.4.

13.2.8 RQ8: What are the main characteristics of the neighbors of power inventors?

To answer this question, we focused on two parameters. The first, called $AvgPatNumNbh_k^{\mathcal{P}}$ represents the average number of patents filed by the neighbors of the power inventors of a country k. Given a country k. This metric can represent an important support to understand how the cooperation with a power inventor is beneficial for a given

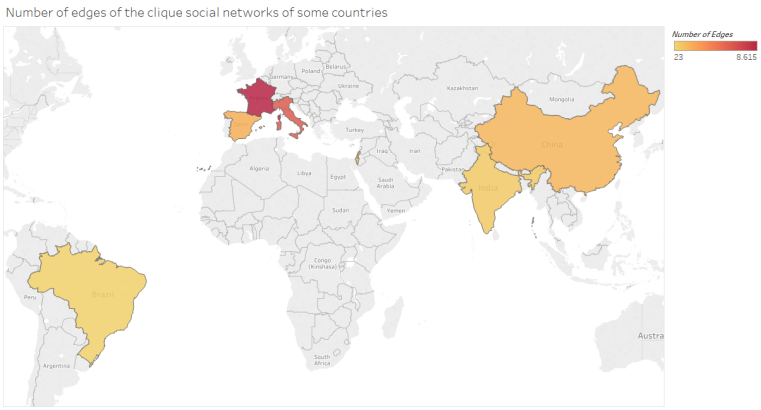


Fig. 13.22. Visualization of the number of edges of the clique social networks of the countries reported in Table 13.4

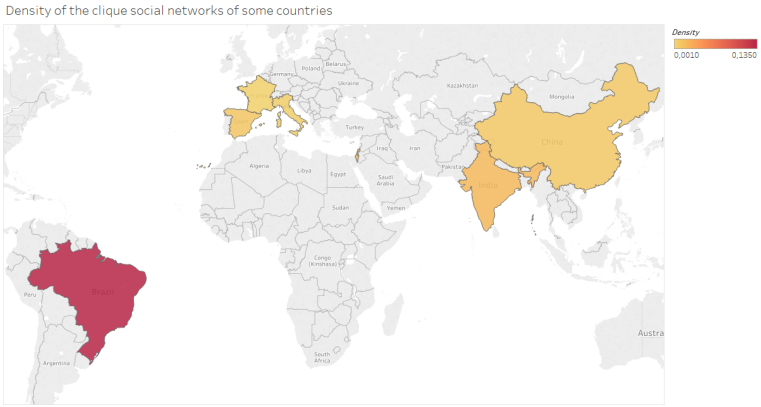


Fig. 13.23. Visualization of the density of the clique social networks of the countries reported in Table 13.4

Country	$AvgPatNumNbh_k^{\mathcal{P}}$	$AvgPatNumNbh_k$	$rPatNumNbh_k$
Brazil	10.979	3.329	3.299
Taiwan	3.437	1.344	2.557
China	6.554	2.680	2.445
India	3.876	2.264	1.712
Italy	7.210	3.397	2.122
Spain	7.668	2.639	2.906
South Africa	2.986	1.466	2.037
Israel	4.675	3.153	1.482

Table 13.5. Average number of patents of the neighbors of a power inventor, of a generic inventor and values of the parameter $rPatNumNbh$

inventor. In fact, high values of this metric indicate that being in the neighborhood of a power inventor stimulates patent filings. In the second column of Table 13.5 we report the obtained values.

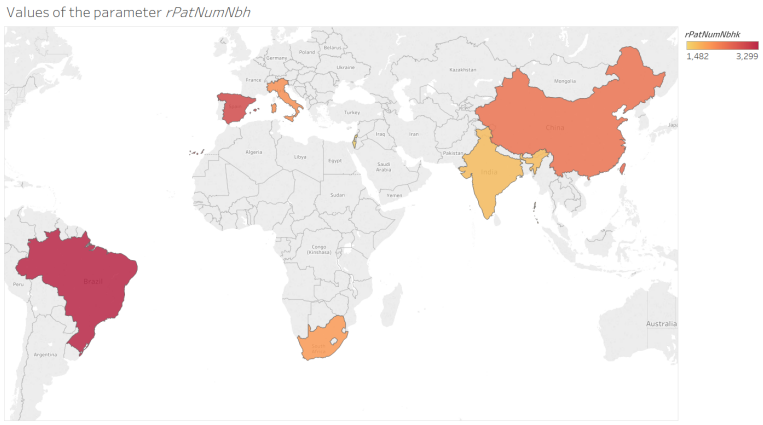


Fig. 13.24. Visualization of the values of $rPatNumNbh_k$ for the countries reported in Table 13.5

To better investigate this issue, we considered the average number $AvgPatNumNbh_k$ of patents filed by the neighbors of the nodes of a given country k. Obtained results are reported in the third column of Table 13.5. The comparison of the second and the third columns of Table 13.5 clearly evidences that being in the neighborhood of a power inventor leads to an increase of the capability of filing patents. To better quantify this intuition, we have defined the following measure:

$$rPatNumNbh_k = \frac{AvgPatNumNbh_k^{\mathcal{P}}}{AvgPatNumNbh_k}.$$

A value of $rPatNumNbh_k$ higher than 1 indicates that belonging to the neighborhood of a power inventor is beneficial for patent filing. The obtained results are reported in the fourth column of Table 13.5. From the analysis of this column, we can observe that obtained values are higher than 1 for all countries. In some countries (e.g., Brazil and Spain), the value of $rPatNumNbh_k$ is higher than or near to 3. This is a clear evidence that the cooperation with power inventors leads to an increase of patent filing. In Figure 13.24, we report a map providing an intuitive visualization of the values of $rPatNumNbh_k$ for the countries reported in Table 13.5.

The second parameter we considered is the average dimension $AvgDimNbh_k^{\mathcal{P}}$ of the neighborhoods of a power inventor. This metric is an indicator of the importance and the centrality of a power inventor. We computed its variation over time for all the countries considered in Table 13.3. For instance, in Figure 13.25, we show the trend of $AvgDimNbh_k^{\mathcal{P}}$ over time for Spain. From the analysis of this figure, we can observe that this parameter remained almost constant for Spain from 1990 to 2003 and, then, had an increase.

Then, we considered the average dimension $AvgDimNbh_k$ of the neighborhoods of a generic node of N_k; in Figure 13.25, we show the trend of this parameter over time

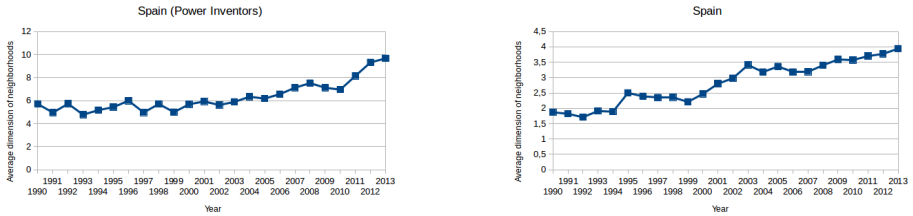


Fig. 13.25. Trend of $AvgDimNbh_k^{\mathcal{P}}$ and $AvgDimNbh_k$ over time for Spain

Country	$rDimNbh$
Austria	2.214
Italy	2.249
Spain	2.454
South Korea	2.360
Russia	2.561
Brazil	1.597
China	2.579
India	2.135
Poland	2.235
South Africa	2.311
Taiwan	2.870
Israel	1.974

Table 13.6. Values of $rDimNbh$ for several countries in the year 2013

for Spain. From the analysis of this figure, we can observe that it tends to increase over time.

Finally, we computed the ratio of the two parameters:

$$rDimNbh_k = \frac{AvgDimNbh_k^{\mathcal{P}}}{AvgDimNbh_k}.$$

When $rDimNbh_k$ is higher than 1, the average dimension of the neighborhoods of power inventors is higher than the corresponding one of generic nodes. In Table 13.6 we report the values of $rDimNbh_k$ for several countries in the year 2013. From the analysis of this table, we can observe that the values obtained for all the countries into consideration are always higher than 1 and range from 1.597 (for Brazil) to 2.870 (for Taiwan). In Figure 13.26, we report a map providing an intuitive visualization of the values of $rDimNbh_k$ for the countries reported in Table 13.6.

13.2.9 RQ9: How are patents distributed against IPC classes?

To answer this question, we computed the distribution of patents against IPC classes for several countries. To give an idea of obtained results, in Figure 13.27, we report the distribution of patents for China and Spain, which represent two extreme cases.

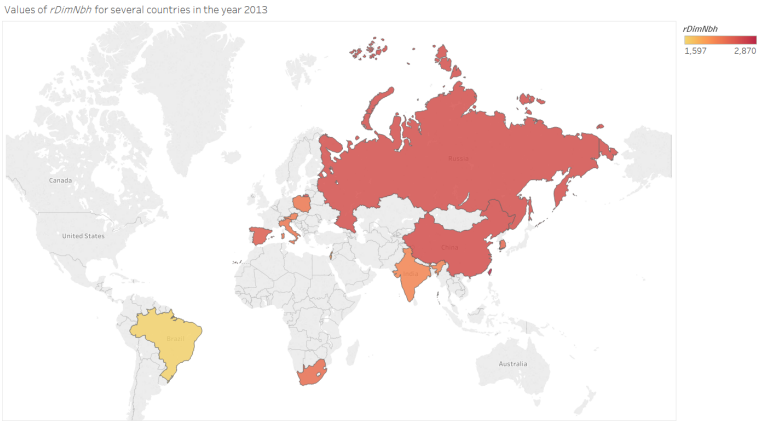


Fig. 13.26. Visualization of the values of $rDimNbh_k$ for the countries reported in Table 13.6

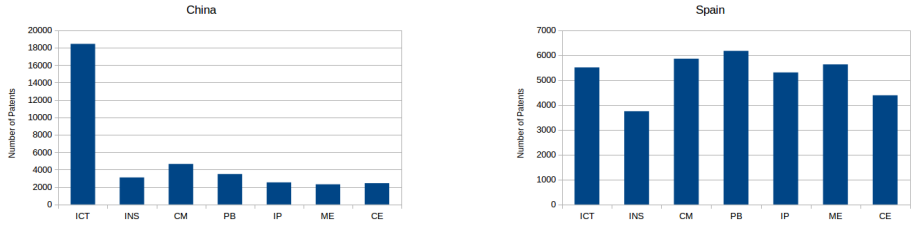


Fig. 13.27. Distribution of patents against IPC classes for China and Spain

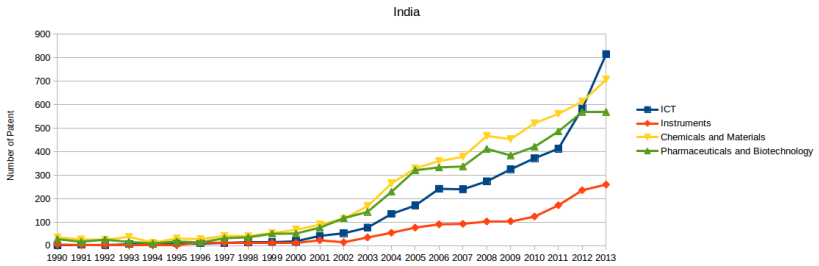


Fig. 13.28. Trend of the distributions of patents against IPC classes for India (Part 1)

In fact, as for China, most of its patents belong to the class "ICT". By contrast, the distribution of Spanish patents is much more uniform.

After this, for several countries, we computed the trend of patent distributions for IPC classes over time. For most of these countries, we observed an increasing trend against time, even if the steepness of the increase is not uniform for the different classes. For instance, in Figures 13.28 and 13.29, we show the trend of patent distributions against IPC classes for India.

To quantify the variety of IPC classes related to the patents of a country k, we computed the corresponding Herfindahl Index $\overline{HI}^*_k$, modified on the basis of what suggested by Hall [190] (see Section 13.2.4). In particular, we carried out this task

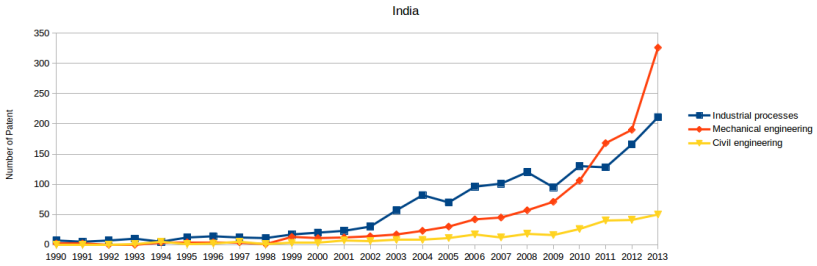


Fig. 13.29. Trend of the distributions of patents against IPC classes for India (Part 2)

Country	$\overline{HI_k^*}$	$\overline{HI_k^*}$ *limited to power inventors*
Brazil	0.156	0.247
Taiwan	0.244	0.350
China	0.294	0.374
Austria	0.151	0.159
Italy	0.151	0.178
Israel	0.196	0.223
Russia	0.157	0.196
Spain	0.146	0.246

Table 13.7. Modified Herfindahl Index concerning the IPC classes of some countries

for several countries. The corresponding results are reported in the second column of Table 13.7.

From the analysis of this table, we can observe that China and Taiwan present the highest values of $\overline{HI_k^*}$, since the corresponding inventors tend to file patents mostly in one IPC classes (in particular, in "ICT" class). In other countries (e.g., Spain) there does not exist a strongly predominant IPC class; for this reason, the corresponding $\overline{HI_k^*}$ is low. In Figure 13.30, we report a map providing an intuitive visualization of the values of the modified Herfindahl Index concerning the IPC classes of the countries reported in Table 13.7.

Finally, we computed $\overline{HI_k^*}$ for the only patents filed by at least one power inventor. Obtained results are shown in the third column of Table 13.7. From the analysis of this column, we can observe that for Brazil, Taiwan and Spain (and, even if to a lesser extent, for China), when passing from generic nodes to power inventors, there is a significant increase of the corresponding Herfindahl Index value. This implies that, in these countries, power inventors, to a much greater extent than the other nodes, tend to focus on only some IPC classes.

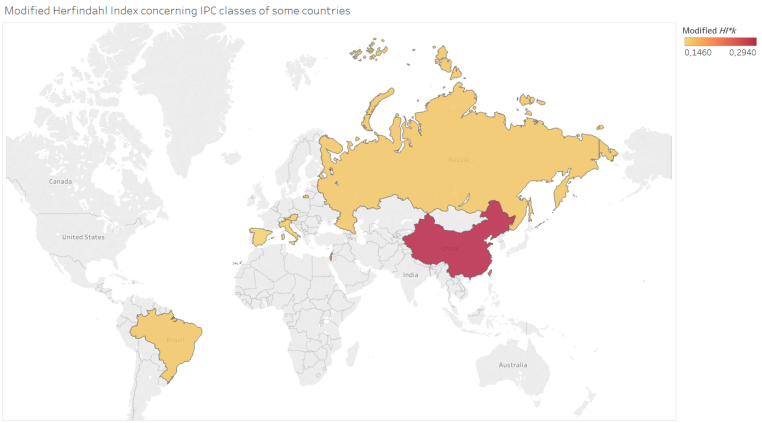


Fig. 13.30. Visualization of the values of the modified Herfindahl Index concerning the IPC classes of the countries reported in Table 13.7

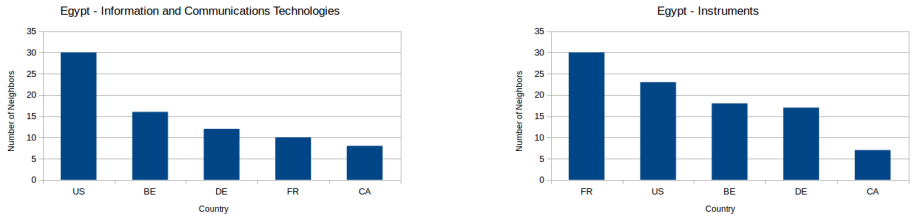


Fig. 13.31. Distribution of the foreign neighbors of the Egyptian inventors for "ICT" and "INS" classes

13.2.10 RQ10: How are foreign collaborations distributed against IPC classes?

To answer this question, given a country k, we computed the distribution of foreign neighbors against IPC classes. Due to space limitations, in the following, we focus on one case that we found extremely interesting, namely Egypt. In Figures 13.31 - 13.34, we show the distribution of the foreign neighbors of the Egyptian inventors for the seven IPC classes. Observe that Egyptian inventors cooperated mainly with: *(i)* US inventors in "ICT", "PB" and "ME" classes; *(ii)* German inventors in the "CM", "IP" and "CE" classes; *(iii)* French inventors in "INS" class.

Another very interesting country from this point of view is Greece. In fact, we found that Greek inventors cooperated mainly with: *(i)* US inventors in "ICT" and "INS" classes; *(ii)* German inventors in "CM" and "IP" classes; *(iii)* Spanish inventors in "PB" class; *(iv)* UK inventors in "ME" and "CE" classes.

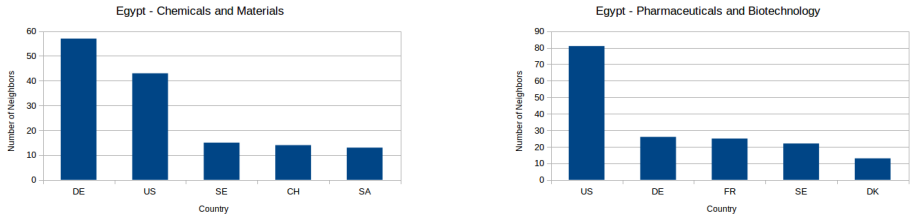


Fig. 13.32. Distribution of the foreign neighbors of the Egyptian inventors for "CM" and "PB" classes

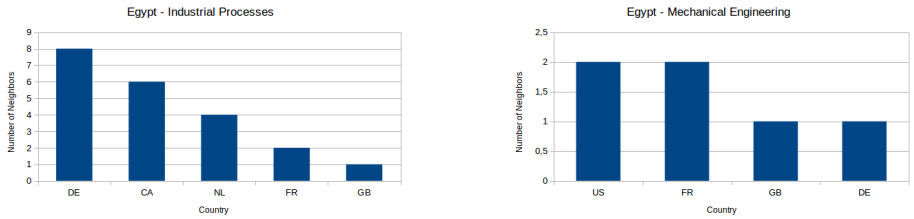


Fig. 13.33. Distribution of the foreign neighbors of the Egyptian inventors for "IP" and "ME" classes

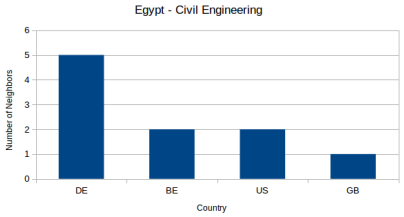


Fig. 13.34. Distribution of the foreign neighbors of the Egyptian inventors for "CE" class

Part V

Closing Remarks

14

Conclusions

In this thesis, we have proposed a unified model and an associated approach to uniformly extract knowledge and face (complex) decision problems in heterogeneous application contexts. We have shown that, thanks to this model and approach, we can manage data coming from several research contexts and we can transpose an approach designed to solve an open problem in one context in such a way as to address other open issues in other contexts. Our model is network-based and our approach is social network analysis-based. We have applied it to four research contexts, namely Biomedical Engineering, Data Lakes, Internet of Things, and Innovation Management. In each of these contexts we have exploited it to address some open problems. In particular, in Biomedical Engineering, we have investigated three neurological disorders, namely Creutzfeldt-Jacob Disease (CJD), Alzheimer Disease (AD) and Childhood Absence Epilepsy (CAE). In Data Lakes, we have adopted it to extract interschema properties and knowledge patterns. In Internet of Things, we have used it to extract knowledge from heterogeneous sensor data streams and to build virtual IoTs in a Multiple IoTs scenario. Finally, in Innovation Management, we have exploited it to extract knowledge concerning patents, their characteristics and their applicants, as well as information about the influence and the scope of a patent on the other ones.

This thesis should not be considered as an ending point. On the contrary, it should be intended as a first step in the attempt to define models and approaches capable of uniformly handling similar issues in very heterogeneous contexts. This could be extremely beneficial for research, because, often, a solution identified in a certain context could be easily transposed in several other ones and, in a scenario characterized by the contamination and interweaving of knowledge, an approach like ours can represent a pioneering attempt in this effort.

As for our future research efforts, first of all we plan to apply our model and approach to several other research contexts, such as Process Mining and Cognitive Computing. Then, we aim at extending our model and approach used for EEG to the analysis of Electrocardiograms in such a way as to investigate heart diseases. Af-

ter this, we plan to extend our approach for patent investigation to the analysis of research papers. Last, but not the least, we aim at extending our approach for extracting knowledge from Data Lakes in such a way as to make it capable of extracting knowledge from images. In this way, it can be well suited to operate in computer vision and robotic vision, which represent two of the hot topics in the current and future ICT research scenario.

References

1. NetworkX. https://networkx.github.io/, 2016.
2. Python. https://www.python.org/, 2016.
3. Thingful: A Search Engine for the Internet of Things. *https://thingful.net/*, 2017.
4. Web Of Science. http://wokinfo.com/, 2017.
5. IPSO Alliance. *https://www.ipso-alliance.org/*, 2019.
6. Mongodb. https://www.mongodb.org/, 2019.
7. The R Project for Statistical Computing. https://www.r-project.org/, 2019.
8. A. Abbas, L. Zhang, and S.U. Khan. A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13, 2014. Elsevier.
9. A. Abbasi. h-Type hybrid centrality measures for weighted networks. *Scientometrics*, 96(2):633–640, 2013. Springer.
10. A. Abbasi, J. Altmann, and L. Hossain. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5 (4):594–607, 2011.
11. A. Abbasi, L. Hossain, and L. Leydesdorff. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6 (3):403–412, 2012.
12. S. Abiteboul and O.M. Duschka. Complexity of answering queries using materialized views. In *Proc. of the International Symposium on Principles of database systems (SIGMOD/PODS'98)*, pages 254–263, Seattle, WA, USA, 1998. ACM.
13. F. J. Acedo, C. Barroso, C. Casanueva, and J. L. Galan. Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies*, 43 (5):957–983, 2006.
14. S. Achard and E. Bullmore. Efficiency and cost of economical brain functional networks. *PLoS Computational Biology*, 3(2):e17, 2007. Public Library of Science.
15. S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E.D. Bullmore. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, 26(1):63–72, 2006. Soc Neuroscience.
16. J. Adams, K. Gurney, D. Hook, and L. Leydesdorff. International collaboration clusters in Africa. *Scientometrics*, 98(1):547–556, 2014. Springer.

17. U. Aguglia, A. Gambardella, E. Le Piane, D. Messina, G. Farnarier, R.L. Oliveri, M. Zappia, and A. Quattrone. Disappearance of periodic sharp wave complexes in Creutzfeldt-Jakob disease. *Neurophysiologie Clinique/Clinical Neurophysiology*, 27(4):277–282, 1997. Elsevier.
18. M. Ahmadlou, A. Adeli, R. Bajo, and H. Adeli. Complexity of functional connectivity networks in mild cognitive impairment subjects during a working memory task. *Clinical Neurophysiology*, 125(4):694–702, 2014. Elsevier.
19. M. Ahmadlou and H. Adeli. Wavelet-synchronization methodology: a new approach for EEG-based diagnosis of ADHD. *Clinical EEG and Neuroscience*, 41(1):1–10, 2010. SAGE Publications.
20. M. Ahmadlou, H. Adeli, and A. Adeli. Improved visibility graph fractality with application for the diagnosis of autism spectrum disorder. *Physica A: Statistical Mechanics and its Applications*, 391(20):4720–4726, 2012. Elsevier.
21. R. K. Ahuja. *Network Flows: Theory, Algorithms, and Applications*. Boston, MA, USA, 2017. Pearson Education.
22. S.A. Akar, S. Kara, F. Latifoğlu, and V. Bilgiç. Analysis of the Complexity Measures in the EEG of Schizophrenia Patients. *International Journal of Neural Systems*, 26(02):1650008, 2016. World Scientific.
23. B.M. Albassuny. Automatic metadata generation applications: a survey study. *International Journal of Metadata, Semantics and Ontologies*, 3(4):260–282, 2008.
24. R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. APS.
25. Z. Aleksovski, M.C.A. Klein, W.T. Kate, and F. van Harmelen. Matching Unstructured Vocabularies Using a Background Ontology. In *Proc. of the International Conference on Knowledge Engineering and Knowledge Management (EKAW'06)*, pages 182–197, Prague, Czech Republic, 2006. Lecture Notes in Computer Science. Springer.
26. A. Algergawy, E. Schallehn, and G. Saake. Improving XML schema matching performance using Pr. *Data & Knowledge Engineering*, 68(8):728–747, 2009. Elsevier.
27. S. P. Algur and P. Bhat. Web Video Object Mining: Expectation Maximization and Density Based Clustering of Web Video Metadata Objects. *International Journal of Information Engineering and Electronic Business*, 8(1):69, 2016. Modern Education and Computer Science Press.
28. F. Ali, S. Islam, D. Kwak, P. Khan, N. Ullah, S. Yoo, and K. Kwak. Type-2 fuzzy ontology–aided recommendation systems for IoT–based healthcare. *Computer Communications*, 2017. Elsevier.
29. T. Alnuaimi, J. Singh, and G. George. Not with my own: long-term effects of cross-country collaboration on subsidiary innovation in emerging economies versus advanced economies. *Journal of Economic Geography*, 12(5):943–968, 2012. Oxford University Press.
30. A. Alserafi, A. Abello, O. Romero, and T. Calders. Towards information profiling: data lake content metadata management. In *Proc. of the International Conference on Data Mining Workshops (ICDMW'16)*, pages 178–185, Barcelona, Spain, 2016. IEEE.

31. J. Alstott, M. Breakspear, P. Hagmann, L. Cammoun, and O. Sporns. Modeling the impact of lesions in the human brain. *PLoS Computational Biology*, 5(6):e1000408, 2009. Public Library of Science.
32. M. Amadeo, C. Campolo, A. Iera, and A. Molinaro. Named data networking for IoT: An architectural perspective. In *Proc. of the European Conference on Networks and Communications (EuCNC'2014)*, pages 1–5, Bologna, Italy, 2014. IEEE.
33. M. Amadeo, C. Campolo, J. Quevedo, D. Corujo, A. Molinaro, A. Iera, R. Aguiar, and A. Vasilakos. Information-centric networking for the internet of things: challenges and opportunities. *IEEE Network*, 30(2):92–100, 2016. IEEE.
34. J. Amezquita-Sanchez, A. Adeli, and H. Adeli. A new methodology for automated diagnosis of mild cognitive impairment (MCI) using magnetoencephalography (MEG). *Behavioural brain research*, 305:174–180, 2016. Elsevier.
35. G. Araniti, A. Orsino, L. Militano, L. Wang, and A. Iera. Context-aware information diffusion for alerting messages in 5G mobile social networks. *IEEE Internet of Things Journal*, 4(2):427–436, 2017. IEEE.
36. T. Arif, R. Ali, and M. Asger. Scientific co-authorship social networks: A case study of computer science scenario in India. *Science*, 52 (12):38–45, 2012.
37. American Psychiatric Association, editor. *Diagnostic and statistical manual of mental disorders (5th ed.)*. 2013.
38. L. Atzori, A. Iera, and G. Morabito. The Internet of Things: A survey. *Computer networks*, 54(15):2787–2805, 2010. Elsevier.
39. L. Atzori, A. Iera, and G. Morabito. SIoT: Giving a social structure to the Internet of Things. *IEEE Communications Letters*, 15(11):1193–1195, 2011. IEEE.
40. L. Atzori, A. Iera, and G. Morabito. From "smart objects" to "social objects": The next evolutionary step of the Internet of Things. *IEEE Communications Magazine*, 52(1):97–105, 2014. IEEE.
41. L. Atzori, A. Iera, and G. Morabito. Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm. *Ad Hoc Networks*, 56:122–140, 2017. Elsevier.
42. L. Atzori, A. Iera, G. Morabito, and M. Nitti. The Social Internet of Things (SIoT)– when social networks meet the Internet of Things: Concept, architecture and network characterization. *Computer networks*, 56(16):3594–3608, 2012. Elsevier.
43. C. Autant-Bernard, S. Chalaye, E. Gagliardini, and S. Usai. European Knowledge Neighbourhood: Knowledge Production in EU Neighbouring Countries and Intensity of the Relationship with EU Countries. *Tijdschrift voor economische en sociale geografie*, 108(1):52–75, 2017. Wiley Online Library.
44. L. Aversano, R. Intonti, C. Quattrocchi, and M. Tortorella. Building a virtual view of heterogeneous data source views. In *Proc. of the International Conference on Software and Data Technologies (ICSOFT'10)*, pages 266–275, Athens, Greece, 2010. INSTICC Press.

45. S. Ayyappan and U. Seneviratne. Electroencephalographic Changes in Sporadic Creutzfeldt–Jakob Disease and Correlation With Clinical Stages: A Retrospective Analysis. *Journal of Clinical Neurophysiology*, 31(6):586–593, 2014. LWW.
46. K. Badar, J.M. Hite, and Y.F. Badir. Examining the relationship of co-authorship network centrality and gender on academic research performance: the case of chemistry researchers in Pakistan. *Scientometrics*, 94 (2):755–775, 2013. Elsevier.
47. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. 1999. Addison Wesley Longman.
48. M. Balconi, S. Breschi, and F. Lissoni. Networks of Inventors and the Location of University Research: An Exploration of Italian Data. In *Proc. of the International Conference "Rethinking Science Policy"*, Brighton, UK, 2002. SPRU.
49. G. Baldassarre, P. Lo Giudice, L. Musarella, and D. Ursino. A paradigm for the cooperation of objects belonging to different IoTs. In *Proc. of the International Database Engineering & Applications Symposium (IDEAS 2018)*, pages 157–164, Villa San Giovanni, Italy, 2018. ACM.
50. G. Baldassarre, P. Lo Giudice, L. Musarella, and D. Ursino. The MIoT paradigm: main features and an "ad-hoc" crawler. *Future Generation Computer Systems*, 92:29–42, 2019. Elsevier.
51. C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Physical review letters*, 88 (17):174102, 2002.
52. A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. American Association for the Advancement of Science.
53. A.L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002. Elsevier.
54. D. Barbera-Tomas, F. Jimenez-Saez, and I. Castello-Molina. Mapping the importance of the real world: The validity of connectivity analysis of patent citations networks. *Research Policy*, 40(3):473–486, 2011. Elsevier.
55. R. Barthwal, S. Misra, and M.S. Obaidat. Finding overlapping communities in a complex network of social linkages and Internet of things. *The Journal of Supercomputing*, 66(3):1749–1772, 2013. Springer.
56. D.S. Bassett, E. Bullmore, B.A. Verchinski, V.S. Mattay, D.R. Weinberger, and A. Meyer-Lindenberg. Hierarchical organization of human cortical networks in health and schizophrenia. *The Journal of Neuroscience*, 28(37):9239–9248, 2008. Soc Neuroscience.
57. D.S. Bassett and E.D. Bullmore. Small-world brain networks. *The Neuroscientist*, 12(6):512–523, 2006. Sage Publications.
58. L. Beltrachini, N. von Ellenrieder, and C.H. Muravchik. General bounds for electrode mislocation on the EEG inverse problem. *Computer Methods and Programs in Biomedicine*, 103(1):1–9, 2011. Elsevier.
59. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59, 1999.

60. S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano. Semantic integration and query of heterogeneous information sources. *Data & Knowledge Engineering*, 36(3):215–249, 2001.
61. J. Bernabé-Moreno, A. Tejeda-Lorente, C. Porcel-Gallego, and E. Herrera-Viedma. Leveraging Localized Social Media Insights for Industry Early Warning Systems. *International Journal of Information Technology & Decision Making*, 17(01):357–385, 2018. World Scientific.
62. P.A. Bernstein, J. Madhavan, and E. Rahm. Generic Schema Matching, Ten Years Later. *Proceedings of the VLDB Endowment*, 4(11):695–701, 2011.
63. C. Besthorn, H. Sattel, C. Geiger-Kabisch, R. Zerfass, and H. Forstl. Parameters of EEG dimensional complexity in Alzheimer's disease. *Electroencephalography and clinical neurophysiology*, 95(2):84–89, 1995. Elsevier.
64. C. Besthorn, R. Zerfass, C. Geiger-Kabisch, H. Sattel, S. Daniel, U. Schreiter-Gasser, and H. Forstl. Discrimination of Alzheimer's disease and normal aging by EEG data. *Electroencephalography and clinical neurophysiology*, 103(2):241–248, 1997. Elsevier.
65. B. Bilalli, A. Abelló, T. Aluja-Banet, and R. Wrembel. Towards intelligent data analysis: the metadata challenge. In *Proc. of the International Conference on Internet of Things and Big Data (IoTBD'16)*, pages 331–338, Rome, Italy, 2016.
66. L. Bing, S. Jiang, W. Lam, Y. Zhang, and S. Jameel. Adaptive Concept Resolution for document representation and its applications in text mining. *Knowledge Based Systems*, 74:1–13, 2015.
67. J. Biskup and D. Embley. Extracting information from heterogeneous information sources using ontologically specified target views. *Information Systems*, 28(3):169–212, 2003. Elsevier.
68. V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. IOP Publishing.
69. Bela Bollobas. *Modern Graph Theory (Graduate Texts in Mathematics)*. Salmon Tower Building, New York City, United States, 2013. Springer.
70. P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972. Taylor & Francis.
71. M. Bordons, J. Aparicio, B. González-Albo, and A.A. Díaz-Faes. The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, 9 (1):135–144, 2015.
72. K. Börner, L. dell'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10 (4):57–67, 2005.
73. L. Bornmann and H. Daniel. What do citation counts measure? A review of studies on citing behavior. *Journal of documentation*, 64(1):45–80, 2008. Emerald Group Publishing Limited.

74. Liotta A. Iacca G. Wrtche H. Bosman, H. Anomaly Detection in Sensor Systems Using Lightweight Machine Learning. *IEEE International Conference on Systems, Man, and Cybernetics*, 81.
75. Liotta A. Iacca G. Wrtche H. Bosman, H. Online extreme learning on fixed- point sensor networks. *IEEE 13th International Conference on Data Mining Workshops*.
76. M. R. Bouadjenek, H. Hacid, and M. Bouzeghoub. Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms. *Information Systems*, 56:1–18, 2016.
77. A. Boukottaya and C. Vanoirbeek. Schema matching for transforming structured documents. In *Proc. of the ACM Symposium on Document Engineering (DocEng'05)*, pages 101–110, Bristol, United Kingdom, 2005. ACM.
78. L. Branstetter, G. Li, and F. Veloso. The rise of the internationalco-invention. *The Changing Frontier: Rethinking Science and Innovation Policy*, pages 135–168, 2015. University of Chicago Press.
79. S. Breschi and F. Lissoni. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, 9(4):439–468, 2009. Oxford University Press.
80. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117, 1998.
81. A.A. Bruzzo, B. Gesierich, M. Santi, C.A. Tassinari, N. Birbaumer, and G. Rubboli. Permutation entropy to detect vigilance changes and preictal states from scalp EEG in epileptic patients. a preliminary study. *Neurological Science*, 29(1):3–9, 2008.
82. F. Buccafurri, V.D. Foti, G. Lax, A. Nocera, and D. Ursino. Bridge Analysis in a Social Internetworking Scenario. *Information Sciences*, 224:1–18, 2013. Elsevier.
83. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Crawling Social Internetworking Systems. In *Proc. of the International Conference on Advances in Social Analysis and Mining (ASONAM 2012)*, pages 505–509, Istanbul, Turkey, 2012. IEEE Computer Society.
84. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Moving from social networks to social internetworking scenarios: The crawling perspective. *Information Sciences*, 256:126–137, 2014. Elsevier.
85. M. Buscema, E. Grossi, M. Capriotti, C. Babiloni, and P. Rossini. The IFAST model allows the prediction of conversion to Alzheimer disease in patients with mild cognitive impairment with high degree of accuracy. *Current Alzheimer Research*, 7(2):173–187, 2010. Bentham Science Publishers.
86. H. Cai, B. Xu, L. Jiang, and A. Vasilakos. IoT-based big data storage systems in cloud computing: perspectives and challenges. *IEEE Internet of Things Journal*, 4(1):75–87, 2017. IEEE.
87. A. Cammarano, F. Michelino, E. Lamberti, and M. Caputo. From social network analysis to business network analysis: Roles and features of companies involved in joint patenting activities. In *Proc. of the International Business Information Management*

Association Conference - Innovation Vision 2020: From Regional Development Sustainability to Global Economic Growth(IBIMA 2015), pages 955–964, Madrid, Spain, 2015.

88. Y. Cao, W.W. Tung, J.B. Gao, V.A. Protopopescu, and L.M. Hively. Detecting dynamical changes in time series using the permutation entropy. *Physical Review E*, 70:046217, 2004.
89. S. Castano, V. De Antonellis, and S. De Capitani di Vimercati. Global viewing of heterogeneous data sources. *IEEE Transactions on Data and Knowledge Engineering*, 13(2):277–297, 2001.
90. F. Cauteruccio, P. Lo Giudice, G. Terracina, and D. Ursino. Applying network analysis for extracting knowledge about environment changes from heterogeneous sensor data streams. In *Multidisciplinary Approaches to Neural Computing*, pages 179–190. 2019. Smart Innovation, Systems and Technologies. Springer.
91. Stamile C. Terracina G. Ursino D. Sappey-Marinier D. Cauteruccio, F. An automated string-based approach to extracting and characterizing White Matter fiber-bundles. *Computers in Biology and Medicine*, 77:64–75.
92. Y. Chang, W.G. Yang, M. Yang, K. Lai, C.Y. Lin, and H.Y. Chang. Locate the technological position by technology redundancy and centralities: Patent citation network perspective. In *Proc. of the International Conference on Management of Engineering and Technology (PICMET 2016)*, pages 1550–1559, Honolulu, HI, USA, 2016. IEEE.
93. C.P. Chen and C.Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275:314–347, 2014.
94. D. Chen, H. Gao, L. Lu, and T. Zhou. Identifying influential nodes in large-scale directed networks: the role of clustering. *PloS one*, 8(10):e77455, 2013. Public Library of Science.
95. J. Chen, N. Zhong, and J. Feng. Developing a Provenance Warehouse for the Systematic Brain Informatics Study. *International Journal of Information Technology & Decision Making*, 16(06):1581–1609, 2017. World Scientific.
96. L. Chen, J. Shao, Z. Yu, J. Sun, F. Wu, and Y. Zhuang. RAISE: A Whole Process Modeling Method for Unstructured Data Management. In *Proc. of the International Conference on Multimedia Big Data (BigMM'15)*, pages 9–12, China National Conference Center, China, 2015. IEEE.
97. L. Chen, M. Tseng, and X. Lian. Development of foundation models for Internet of Things. *Frontiers of Computer Science in China*, 4(3):376–385, 2010. Springer.
98. X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proc. of the International Conference on Computer Vision (ICCV'13)*, pages 1409–1416, Darling Harbour, Sydney, 2013. IEEE.
99. Y. Chen, C. Ding, J. Hu, R. Chen, P. Hui, and X. Fu. Building and analyzing a global co-authorship network using google scholar data. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1219–1224, Perth, Australia, 2017. International World Wide Web Conferences Steering Committee.

100. Y. Chen, W. Wang, and Z. Liu. Keyword-based search and exploration on databases. In *Proc. of the International Conference on Data Engineering (ICDE'11)*, pages 1380–1383, Hannover, Germany, 2011. IEEE.

101. Y. Chen, Z. Zhen, H. Yu, and J. Xu. Application of Fault Tree Analysis and Fuzzy Neural Networks to Fault Diagnosis in the Internet of Things (IoT) for Aquaculture. *Sensors*, 17(1):153, 2017. Multidisciplinary Digital Publishing Institute.

102. J. Chiang, Z. Wang, and M. McKeown. A generalized multivariate autoregressive (GmAR)-based approach for EEG source connectivity analysis. *IEEE Transactions on Signal Processing*, 60(1):453–465, 2012. IEEE.

103. Z. Chinchilla-Rodriguez, A. Ferligoj, S. Miguel, L. Kronegger, and F. de Moya-Anegón. Blockmodeling of co-authorship networks in library and information science in Argentina: a case study. *Scientometrics*, 93 (3):699–717, 2012.

104. B. Christophe, V. Verdot, and V. Toubian. Searching the 'web of things'. In *Proc. of the International Conference on Semantic Computing (ICSC'2011)*, pages 308–315, Palo Alto, CA, USA, 2011. IEEE.

105. D. Chyzhyk, M. Graña, D. Öngür, and A.K. Shinn. Discrimination of schizophrenia auditory hallucinators by machine learning of resting-state functional MRI. *International Journal of Neural Systems*, 25(03):1550007, 2015. World Scientific.

106. A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review Part E*, 70(6):066111, 2004.

107. A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

108. M. Coffano and G. Tarasconi. CRIOS - Patstat Database: Sources, Contents and Access Rules. *Center for Research on Innovation, Organization and Strategy, CRIOS Working Paper*, 2014.

109. L. Cooke and H. Hall. Facets of DREaM: A social network analysis exploring network development in the UK LIS research community. *Journal of Documentation*, 69(6):786–806, 2013. Emerald Group Publishing Limited.

110. A. Corbellini, C. Mateos, A. Zunino, D. Godoy, and S.N. Schiaffino. Persisting big-data: The NoSQL landscape. *Information Systems*, 63:1–23, 2017. Elsevier.

111. L. Costa, F. Rodrigues, G. Travieso, and P. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242, 2007. Taylor & Francis.

112. D. Cui, W. Pu, J. Liu, Z. Bian, Q. Li, L. Wang, and G. Gu. A new EEG synchronization strength analysis method: S-estimator based normalized weighted-permutation mutual information. *Neural Networks*, 82:30–38, 2016. Elsevier.

113. Tropea-M. Fazio P. Marano S. Curia, V. Complex networks: Study and performance evaluation with hybrid model for Wireless Sensor Networks. *In Electrical and Computer Engineering*.

114. Papadimitriou-A. Katsaros D. Manolopoulos Y. Cuzzocrea, A. Edge betweenness centrality: A novel algorithm for QoS-based topology control over wireless sensor networks. *Journal of Network and Computer Applications*, 35:1210–1217, 2017. Elsevier.

115. B. Czigler, D. Csikós, Z. Hidasi, Z. Gaál, E. Csibri, E. Kiss, P. Salacz, and M. Molnár. Quantitative EEG in early Alzheimer's disease patientsópower spectrum and complexity features. *International Journal of Psychophysiology*, 68(1):75–80, 2008. Elsevier.
116. A. Dass, C. Aksoy, A. Dimitriou, and D. Theodoratos. Relaxation of keyword pattern graphs on RDF Data. *Journal of Web Engineering*, 16(5-6):363–398, 2017. Rinton Press, Incorporated.
117. J. Dauwels, F. Vialatte, and A. Cichocki. Diagnosis of Alzheimer's disease from EEG signals: where are we standing? *Current Alzheimer Research*, 7(6):487–505, 2010. Bentham Science Publishers.
118. J. Dauwels, F. Vialatte, and A. Cichocki. On the early diagnosis of Alzheimer's disease from EEG signals: A mini-review. In *Advances in Cognitive Neurodynamics (II)*, pages 709–716. Springer, 2011.
119. J. Dauwels, F. Vialatte, T. Musha, and A. Cichocki. A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG. *NeuroImage*, 49(1):668–693, 2010. Elsevier.
120. C. Davatzikos, P. Bhatt, L.M. Shaw, K.N. Batmanghelich, and J.Q. Trojanowski. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of aging*, 32(12):2322–e19, 2011. Elsevier.
121. R. De. Optimal conditions for innovation: Firm-level evidence from Kenya and Uganda. In *Proc. of the Annual African Economic Conference*, Addis Abeba, Ethiopia, 2014.
122. W. de Haan, Y.A. Pijnenburg, R.L. Strijers, Y. van der Made, W. van der Flier, P. Scheltens, and C.J. Stam. Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory. *BMC neuroscience*, 10(1):1, 2009. BioMed Central.
123. P. De Meo, G. Quattrone, G. Terracina, and D. Ursino. "Almost automatic" and semantic integration of XML Schemas at various "severity levels". In *Proc. of the International Conference on Cooperative Information Systems (CoopIS 2003)*, pages 4–21, Taormina, Italy, 2003. Lecture Notes in Computer Science, Springer.
124. P. De Meo, G. Quattrone, G. Terracina, and D. Ursino. Integration of XML Schemas at various "severity" levels. *Information Systems*, 31(6):397–434, 2006.
125. T. Dehdarirad and S. Nasini. Research impact in co-authorship networks: a two-mode analysis. *Journal of Informetrics*, 11(2):371–388, 2017. Elsevier.
126. A. Delorme and S. Makeig. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including Independent Component Analysis. *Journal of Neuroscience Methods*, 134:9–21, 2004.
127. N. Deo. *Graph Theory with Applications to Engineering and Computer Science*. Mineola, New York, United States, 2016. Dover Publications.
128. F. Di Tria, E. Lefons, and F. Tangorra. Cost-benefit analysis of data warehouse design methodologies. *Information Systems*, 63:47–62, 2017. Elsevier.
129. C. Diamantini, P. Lo Giudice, L. Musarella, D. Potena, E. Storti, and D. Ursino. An approach to extracting thematic views from highly heterogeneous sources of a data

lake. In *Atti del Ventiseiesimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati (SEBD'18)*, Castellaneta Marina (TA), Italy, 2018.

130. D. Diamantini, D. Potena, and E. Storti. A virtual mart for knowledge discovery in databases. *Information Systems Frontiers*, 15(3):447–463, 2013. Springer.
131. S. Distefano, G. Merlino, and A. Puliafito. Enabling the cloud of things. In *Proc. of the International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS'2012)*, pages 858–863, Taichung, Taiwan, 2012. IEEE.
132. M. Diykh, Y. Li, and P. Wen. Classify epileptic EEG signals using weighted complex networks based community structure detection. *Expert Systems with Applications*, 90:87–100, 2017. Elsevier.
133. H. Dong, F. Hussain, and E. Chang. A framework for discovering and classifying ubiquitous services in digital health ecosystems. *Journal of Computer and System Sciences*, 77(4):687–704, 2011. Elsevier.
134. S. Dorogovtsev, A. Goltsev, and J. Mendes. K-core organization of complex networks. *Physical review letters*, 96(4):040601, 2006. APS.
135. R. dos Santos Mello, S. Castano, and C.A. Heuser. A method for the unification of XML schemata. *Information & Software Technology*, 44(4):241–249, 2002. Elsevier.
136. A. Dulamea and E. Solomon. Role of the biomarkers for the diagnosis of Creutzfeldt-Jakob disease. *Journal of medicine and life*, 9(2):211, 2016. Carol Davila-University Press.
137. J. Duun-Henriksen, T.W. Kjaer, R.E. Madsen, L.S. Remvig, C.E. Thomsen, and H.B. Sorensen. Channel selection for automatic seizure detection. *Clinical Neurophysiology*, 123(1):84–92, 2012.
138. J. Duun-Henriksen, R.E. Madsen, L.S. Remvig, C.E. Thomsen, H.B. Sorensen, and T.W. Kjaer. Automatic detection of childhood absence epilepsy seizures: toward a monitoring device. *Pediatric Neurology*, 46(5):287–292, 2012.
139. B.D.P.F. e Fonseca, R. Sampaio, M.V. de Araujo Fonseca, and F. Zicker. Co-authorship network analysis in health research: method and potential use. *Health Research Policy and Systems*, 14(1):34, 2016. BioMed Central.
140. E. Eggenberger and D. Murman. Prion Diseases and Other Rapidly Progressive Dementias. *Neurodegeneration*, page 126, 2017. John Wiley & Sons.
141. O. Ejermo and C. Karlsson. Interregional inventor networks as studied by patent coinventorships. *Research Policy*, 35(3):412–430, 2006. Elsevier.
142. P. Ellis, G. Hepburn, and C. Oppenhein. Studies on patent citation networks. *Journal of documentation*, 34(1):12–20, 1978. MCB UP Ltd.
143. H. Elmeleegy, M. Ouzzani, and A.K. Elmagarmid. Usage-Based Schema Matching. In *Proc. of the International Conference on Data Engineering (ICDE'08)*, pages 20–29, Cancún, México, 2008. IEEE.
144. P. Erdi and P. Bruck. Patent Citation Network Analysis: Ranking: from web pages to patents. In *Proc. of the International Conference on Artificial Neural Networks, (ICANN 2016)*, page 544, Barcelona, Spain, 2016. Springer Verlag.

145. E. Estrada and J. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005. APS.

146. A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A.Y. Zomaya, S. Foufou, and A. Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279, 2014. IEEE.

147. W. Fan, X. Wang, and Y. Wu. Answering pattern queries using views. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):326–341, 2016. IEEE.

148. H. Fang. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *Proc. of the International Conference on Cyber Technology in Automation (CYBER'15)*, pages 820–824, Shenyang, China, 2015. IEEE.

149. M. Farid, A. Roatis, I.F. Ilyas, H. Hoffman, and X. Chu. CLAMS: bringing quality to Data Lakes. In *Proc. of the International Conference on Management of Data (SIGMOD/PODS'16)*, pages 2089–2092, San Francisco, CA, USA, 2016. ACM.

150. I. Farris, R. Girau, L. Militano, M. Nitti, L. Atzori, A. Iera, and G. Morabito. Social virtual objects in the edge cloud. *IEEE Cloud Computing*, 2(6):20–28, 2015. IEEE.

151. A. Farrugia, R. Claxton, and S. Thompson. Towards social network analytics for understanding and managing enterprise data lakes. In *Proc. of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM'16)*, pages 1213–1220, San Francisco, CA, USA, 2016. IEEE.

152. F. Feng and W.B. Croft. Probabilistic techniques for phrase extraction. *Information Processing & Management*, 37(2):199–220, 2001.

153. M. Ferrara, D. Fosso, D. Lanatà, R. Mavilia, and D. Ursino. A Social Network Analysis based approach to extracting knowledge patterns about innovation geography from patent databases. *International Journal of Data Mining, Modelling and Management*, 10(1):23–71, 2018. Inderscience.

154. C. Fetzer, P. Felber, E. Riviere, V. Schiavoni, and P. Sutra. Unicrawl: A practical geographically distributed web crawler. In *Proc. of the International Conference on Cloud Computing (CLOUD'2015)*, pages 389–396, New York, NY, USA, 2015. IEEE.

155. T. Foley, H. Hagen, and G. Nielson. Visualizing and modeling unstructured data. *The Visual Computer*, 9(8):439–449, 1993. Springer.

156. R. Fontana, A. Nuvolari, H. Shimizu, and A. Vezzulli. Reassessing patent propensity: evidence from a data-set of R&D awards, 1977-2004. *Research Policy*, 42(10):1780–1792, 2013. Elsevier.

157. R. Fontana, A. Nuvolari, and B. Verspagen. Mapping technological trajectories as patent citation networks. An application to data communication standards. *Economics of Innovation and New Technology*, 18(4):311–336, 2009. Taylor & Francis.

158. E. Forti, C. Franzoni, and M. Sobrero. Bridges or isolates? Investigating the social networks of academic inventors. *Research Policy*, 42:1378–1388, September 2013. Elsevier.

159. Guerrieri A. OHare G. Ruzzelli A. Fortino, G. A flexible building management framework based on wireless sensor and actuator networks. *Journal of Network and Computer Applications*, 35:1934–1952, 2017. Elsevier.

160. F.J. Fraga, T.H. Falk, P.A. Kanda, and R. Anghinah. Characterizing Alzheimerós disease severity via resting-awake EEG amplitude modulation analysis. *PloS one*, 8(8):e72240, 2013. Public Library of Science.

161. L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. JSTOR.

162. L. C. Freeman. Centrality in Social Networks Conceptual and Clarification. *Social Networks*, 1(3):215–239, 1979. Elsevier.

163. J.L. Furman, M.K. Kyle, I.M. Cockburn, and R. Henderson. Public & private spillovers, location and the productivity of pharmaceutical research. Technical report, National Bureau of Economic Research, 2006.

164. E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 1606–1611, Hyderabad, India, 2007.

165. M. Ganzha, M. Paprzycki, W. Pawłowski, P. Szmeja, and K. Wasielewska. Anomaly detection in wireless sensor networks: A survey. *Journal of Network and Computer Applications*, 34:1302–1325, 2017. Elsevier.

166. M. Ganzha, M. Paprzycki, W. Pawłowski, P. Szmeja, and K. Wasielewska. Semantic interoperability in the Internet of Things: An overview from the INTER-IoT perspective. *Journal of Network and Computer Applications*, 81:111–124, 2017. Elsevier.

167. S. Gasparini, E. Ferlazzo, D. Branca, A. Labate, V. Cianci, M.A. Latella, and U. Aguglia. Teaching NeuroImages: Pseudohypertrophic cerebral cortex in end-stage Creutzfeldt-Jakob disease. *Neurology*, 80(2):e21–e21, 2013. AAN Enterprises.

168. R. Gaur and D.K. Sharma. Review of ontology based focused crawling approaches. In *Proc. of the International Conference on Soft Computing Techniques for Engineering and Technology (ICSCTET'2014)*, pages 1–4, Nainital, India, 2014. IEEE.

169. P. Lo Giudice, D. Ursino, N. Mammone, F.C. Morabito, U. Aguglia, V. Cianci, E. Ferlazzo, and S. Gasparini. A network analysis based approach to characterizing Periodic Sharp Wave Complexes in electroencephalograms of patients with sporadic CJD. *International Journal of Medical Informatics*, 121:19–29, 2019. Elsevier.

170. E. Giuliani, A. Martinelli, and R. Rabellotti. Is Co-Invention Expediting Technological Catch Up? A Study of Collaboration between Emerging Country Firms and EU inventors. *World Development*, 77:192–205, 2016. Elsevier.

171. M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *Proc. of the International Conference on Computer Communications (INFOCOM'10)*, pages 1–9, San Diego, CA, USA, 2010. IEEE.

172. Fonseca R. Jamieson K. Kazandjieva M. Moss-D. Levis P. Gnawali, O. An Efficient, Robust, and Reliable Collection Tree Protocol for Wireless Sensor Networks. *ACM Transactions on Sensors Networks*, 10:1–49.

173. M. Goedhuys. Learning, product innovation, and firm heterogeneity in developing countries; Evidence from Tanzania. *Industrial and Corporate Change*, 16(2):269–292, 2007. Oxford University Press.

174. K. Golenberg, B. Kimelfeld, and Y. Sagiv. Keyword proximity search in complex data graphs. In *Proc. of the International Conference on Management of data (SIGMOD-/PODS'08)*, pages 927–940, Vancouver, Canada, 2008. ACM.

175. J. Gotman, J.R. Ives, and P. Gloor. Frequency content of EEG and EMG at seizure onset: possibility of removal of textscEMG artefact by digital filtering. *Electroencephalography and clinical neurophysiology*, 52(6):626–639, 1981. Elsevier.

176. A.A. Gouw, A.M. Alsema, B.M. Tijms, A. Borta, P. Scheltens, C.J. Stam, and W.M. van der Flier. EEG spectral analysis as a putative early prognostic biomarker in nondemented, amyloid positive subjects. *Neurobiology of aging*, 57:133–142, 2017. Elsevier.

177. R.M. Grant. Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science*, 7(4):375–387, 1996. INFORMS.

178. Z. Griliches. Introduction to "output measurement in the service sectors". In *Output measurement in the service sectors*, pages 1–22. 1992. University of Chicago Press.

179. J. L. Gross and J. Yellen. *Graph Theory and Its Applications*. New York, United States, 2005. Chapman and Hall/CRC.

180. J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645–1660, 2013. Elsevier.

181. D. Guellec and B. van Pottelsberghe de la Potterie. The internationalisation of technology analysed with patent data. *Research Policy*, 30(8):1253–1266, 2001. Elsevier.

182. D. Guinard, M. Fischer, and V. Trifa. Sharing using social networks in a composable web of things. In *Proc. of the International Conference on Pervasive Computing and Communications (PERCOM 2010)*, pages 702–707, Mannheim, Germany, 2010. IEEE.

183. D. Guinard, V. Trifa, F. Mattern, and E. Wilde. From the internet of things to the web of things: Resource-oriented architecture and best practices. *Architecting the Internet of Things*, pages 97–129, 2011. Springer.

184. D. Guinard, V. Trifa, and E. Wilde. Architecting a mashable open world wide web of things. *Technical Report of the Institute for Pervasive Computing, ETH Zürich, Zürich, Switzerland*, 663, 2010.

185. D. Gutiérrez and D.I.Escalona-Vargas. EEG data classification through signal spatial redistribution and optimized linear discriminants. *Computer Methods and Programs in Biomedicine*, 97(1):39–47, 2010. Elsevier.

186. P. Hage and F. Harary. Eccentricity and centrality in networks. *Social networks*, 17(1):57–63, 1995. Elsevier.

187. P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C.J. Honey, J.V. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLoS Computational Biology*, 6(7):e159, 2008. Public Library of Science.

188. R. Hai, S. Geisler, and C. Quix. Constance: An intelligent data lake system. In *Proc. of the International Conference on Management of Data (SIGMOD 2016)*, pages 2097–2100, San Francisco, CA, USA, 2016. ACM.

189. A. Halevy. Answering queries using views: A survey. *The VLDB Journal*, 10(4):270–294, 2001. Springer.

190. B.H. Hall. A Note on the Bias in Herfindahl-Type Measures Based on Count Data. *Revue D'Economie Industrielle*, 110:149–159, 2005. Editions Techniques et Economiques, CNRS.

191. J. Han and M. Kamber. *Data Mining: Concepts and Techniques - Second Edition*. Morgan Kaufmann notes, 2006.

192. S. Han, L. Zou, J.X. Yu, and D. Zhao. Keyword Search on RDF Graphs-A Query Graph Assembly Approach. In *Proc. of the International Conference on Information and Knowledge Management (CIKM'17)*, pages 227–236, Singapore, Singapore, 2017. ACM.

193. R. Hanneman and M. Riddle. *Introduction to social network methods*. `http://faculty.ucr.edu/~hanneman/nettext/` , 2005. University of California, Riverside.

194. S.M. Harding, W.B. Croft, and C. Weir. Probabilistic retrieval of ocr degraded text using n-grams. In *Proc. of the International Conference on Theory and Practice of Digital Libraries (ECDL'97)*, pages 345–359, Pisa, Italy, 1997. Springer.

195. F. Hatz, M. Hardmeier, N. Benz, M. Ehrensperger, U. Gschwandtner, S. Rüegg, C. Schindler, A.U. Monsch, and P. Fuhr. Microstate connectivity alterations in patients with early Alzheimers disease. *Alzheimer's research & therapy*, 7(1):78, 2015. BioMed Central.

196. H. He, H. Wang, J. Yang, and P.S. Yu. BLINKS: ranked keyword searches on graphs. In *Proc. of the International Conference on Management of Data (SIGMOD/PODS'07)*, pages 305–316, Beijing, China, 2007. ACM.

197. Y. He, Z.J. Chen, and A.C. Evans. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cerebral Cortex*, 17(10):2407–2419, 2007. Oxford University Press.

198. L. Hebert, J. Weuve, P. Scherr, and D. Evans. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783, 2013. AAN Enterprises.

199. V.A. Hill. Collaboration in an Academic Setting: Does the Network Structure Matter? *Center for the Computational Analysis of Social and Organizational Systems (CASOS) technical report*, 2008.

200. P. Hingley and S. Bas. Numbers and sizes of applicants at the European Patent Office. *World Patent Information*, 31(4):285–298, 2009. Elsevier.

201. J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005. National Academy of Sciences.

202. T.J. Hirschauer, H. Adeli, and J.A. Buford. Computer-aided diagnosis of Parkinson's disease using enhanced probabilistic neural network. *Journal of Medical Systems*, 39(11):179, 2015. Springer.

203. A.O. Hirschman. The paternity of an index. *The American Economic Review*, 54(5):761–762, 1964.

204. et al. Ho, J.-W. Distributed detection of replica node attacks with group deployment knowledge in wireless sensor networks. *Ad Hoc Networks*.

205. L. Holmquist, F. Mattern, B. Schiele, P. Alahuhta, M. Beigl, and H. Gellersen. Smart-its friends: A technique for users to easily establish connections between smart artefacts. In *Proc. of the International Conference on Ubiquitous Computing (Ubicomp'2001)*, pages 116–122, Atlanta, GA, USA, 2001. Springer.

206. C.J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J.P. Thiran, R. Meuli, and P. Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proc. of the National Academy of Sciences*, 106(6):2035–2040, 2009. National Academy of Sciences.

207. R. Hornero, D. Abásolo, J. Escudero, and C. Gómez. Nonlinear analysis of electroencephalogram and magnetoencephalogram recordings in patients with Alzheimer's disease. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1887):317–336, 2009. The Royal Society.

208. C.C. Hsueh and C.C. Wang. The Use of Social Network Analysis in Knowledge Diffusion Research from Patent Data. In *Proc. of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2009)*, pages 393–398, Athens, Greece, 2009. IEEE Computer Society.

209. J. Hu and Y. Zhang. Structure and patterns of cross-national Big Data research collaborations. *Journal of Documentation*, 73(6):1119–1136, 2017. Emerald Publishing Limited.

210. J. Huenteler, J. Ossenbrink, T. Schmidt, and V. Hoffmann. How a product's design hierarchy shapes the evolution of technological knowledge – Evidence from patent-citation networks in wind power. *Research Policy*, 45(6):1195–1217, 2016. Elsevier.

211. J. Huenteler, T. Schmidt, J. Ossenbrink, and V. Hoffmann. Technology life-cycles in the energy sector – Technological characteristics and the role of deployment for innovation. *Technological Forecasting and Social Change*, 104:102–121, 2016. Elsevier.

212. N. Hummon and P. Dereian. Connectivity in a citation network: The development of DNA theory. *Social networks*, 11(1):39–63, 1989. Elsevier.

213. N.Q.V. Hung, N.T. Tam, V.T. Chau, T.K. Wijaya, Z. Miklós, K. Aberer, A. Gal, and M. Weidlich. SMART: A tool for analyzing and reconciling schema matching networks. In *Proc. of the International Conference on Data Engineering (ICDE'15)*, pages 1488–1491, Seoul, South Korea, 2015. IEEE.

214. S. Hung and A. Wang. Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics*, 82(1):121–134, 2010. Springer.

215. I. Ishaq, D. Carels, G. Teklemariam, J. Hoebeke, F. Abeele, E. Poorter, I. Moerman, and P. Demeester. IETF standardization in the field of the Internet of Things (IoT): a survey. *Journal of Sensor and Actuator Networks*, 2(2):235–287, 2013. Multidisciplinary Digital Publishing Institute.

216. A. Jain. Betweenness centrality based connectivity aware routing algorithm for prolonging network lifetime in wireless sensor networks. *Wireless Networks*.

217. S. Jain and S. Tanwani. Schema matching technique for heterogeneous web database. In *Proc. of the International Conference on Reliability (ICRITO'15)*, pages 1–6, Noida, India, 2015. IEEE.

218. M. Jalili. Graph theoretical analysis of Alzheimer's disease: Discrimination of AD patients from healthy subjects. *Information Sciences*, 384:145–156, 2017. Elsevier.

219. F. Jebbor and L. Benhlima. Overview of knowledge extraction techniques in five question-answering systems. In *Proc. of the International Conference on Intelligent Systems: Theories and Applications (SITA'14)*, pages 1–8, Rabat, Morocco, 2014. IEEE.

220. J. Jeong. EEG dynamics in patients with Alzheimer's disease. *Clinical neurophysiology*, 115(7):1490–1505, 2004. Elsevier.

221. S. Jiang, L. Bing, B. Sun, Y. Zhang, and W. Lam. Ontology enhancement and concept granularity learning: keeping yourself current and adaptive. In *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD'11)*, pages 1244–1252, San Diego, CA, USA, 2011. ACM.

222. Q. Jing, A. Vasilakos, J. Wan, J. Lu, and D. Qiu. Security of the Internet of Things: perspectives and challenges. *Wireless Networks*, 20(8):2481–2501, 2014. Springer.

223. K.Y. Jung, D.W. Seo, D.L. Na, C.S. Chung, I.K. Lee, K. Oh, C.H. Im, and H.K. Jung. Source localization of periodic sharp wave complexes using independent component analysis in sporadic Creutzfeldt–Jakob disease. *Brain Research*, 1143:228–237, 2007. Elsevier.

224. A.B. Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.

225. N. Kannathal, U.R. Acharya, C.M. Lim, and P.K. Sadasivan. Characterization of EEG - a comparative study. *Computer Methods and Programs in Biomedicine*, 80(1):17–23, 2005. Elsevier.

226. S. Kapidakis. Rating quality in metadata harvesting. In *Proc. of the International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'15)*, pages 65:1–65:8, New York, NY, USA, 2015. ACM.

227. M. Kargar and A. An. Keyword search in graphs: Finding r-cliques. *Proceedings of the VLDB Endowment*, 4(10):681–692, 2011. VLDB Endowment.

228. M.D. Karim, M. Cochez, O.D. Beyan, C.F. Ahmed, and S. Decker. Mining maximal frequent patterns in transactional databases and dynamic data streams: A spark-based approach. *Information Sciences*, 432:278–300, 2018.

229. S. Karnouskos. Smart houses in the smart grid and the search for value-added services in the cloud of things era. In *Proc. of the International Conference on Industrial*

Technology (ICIT'2013), pages 2016–2021, Cape Town, Western Cape, South Africa, 2013. IEEE.

230. V. Karyotis, K. Tsitseklis, K. Sotiropoulos, and S. Papavassiliou. Big Data Clustering via Community Detection and Hyperbolic Network Embedding in IoT Applications. *Sensors*, 18(4):1205, 2018. Multidisciplinary Digital Publishing Institute.
231. N. Kasabov and E. Capecci. Spiking neural network methodology for modelling, classification and understanding of EEG spatio-temporal data measuring cognitive processes. *Information Sciences*, 294:565–575, 2015. Elsevier.
232. L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. Springer.
233. S. Kazi, Q. Rajput, and S. Khoja. Study of Evolving Co-Authorship Network: Identification of Growth Patterns of Collaboration Using SNA Measures. In *IEEE 11th International Conference on Semantic Computing*, pages 488–493, San Diego, CA, USA, 2017. IEEE.
234. H. Kim, P. Howland, and H. Park. Dimension Reduction in Text Classification with Support Vector Machines. *Journal of Machine Learning Research*, 6:37–53, 2005.
235. H.K. Kim, H. Kim, and S. Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352, 2017.
236. J. Kim, S. J. Lee, and G. Marschke. International knowledge flows: evidence from an inventor-firm matched data set. In *Science and Engineering Careers in the United States: An Analysis of Markets and Employment*, pages 321–348. 2009. University of Chicago Press.
237. J. Kim and C. Perez. Co-authorship network analysis in industrial ecology research community. *Journal of Industrial Ecology*, 19 (2):222–235, 2015.
238. C. Kiss and M. Bichler. Identification of influencersómeasuring influence in customer networks. *Decision Support Systems*, 46(1):233–253, 2008. Elsevier.
239. J.K. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. ACM.
240. M. Knyazeva, M. Jalili, A. Brioschi, I. Bourquin, E. Fornari, M. Hasler, R. Meuli, P. Maeder, and J. Ghika. Topography of EEG multivariate phase synchronization in early Alzheimer's disease. *Neurobiology of aging*, 31(7):1132–1144, 2010. Elsevier.
241. G. Kondrak. N-gram similarity and distance. In *String Processing and Information Retrieval*, pages 115–126, 2005. Springer.
242. M. Koppert, S. Kalitzin, D. Velis, F. Lopes Da Silva, and M.A. Viergever. Preventive and Abortive Strategies for Stimulation Based Control of Epilepsy: A Computational Model Study. *International Journal of Neural Systems*, 26(08):1650028, 2016. World Scientific.
243. D. Koschützki, K.A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski. Centrality indices. In *Network analysis*, pages 16–61. 2005. Springer.

244. A.M. Kowshalya and M.L. Valarmathi. Community Detection in the Social Internet of Things Based on Movement, Preference and Social Similarity. *Studies in Informatics and Control*, 25(4):499–506, 2016. National Institute for R&D in Informatics.

245. M. Kranz, L. Roalter, and F. Michahelles. Things that Twitter: social networks and the Internet of Things. In *Proc. of the International Workshop on Pervasive Computing (Pervasive 2010)*, pages 1–10, Helsinki, Finland, 2010.

246. H. Kretschmer and T. Kretschmer. A new centrality measure for social network analysis applicable to bibliometric and webometric data. *Collnet Journal of Scientometrics and Information Management*, 1(1):1–7, 2007. Taylor & Francis.

247. B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *Proc. of the First Workshop on Online Social Networks (WOSN'2008)*, pages 19–24, Seattle, WA, USA, 2008.

248. H. Kumarage, I. Khalil, Z. Tari, and A. Zomaya. Distributed anomaly detection for industrial wireless sensor networks based on fuzzy data modelling. *Journal of Parallel and Distributed Computing*, 73(6):790–806, 2013. Elsevier.

249. M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS (Breadth First Search). In *Proc. of the International Teletraffic Congress (ITC'2010)*, pages 1–8, Amsterdam, The Netherlands, 2010. IEEE.

250. D. Labate, F. La Foresta, G. Morabito, I. Palamara, and F.C. Morabito. Entropic measures of EEG complexity in Alzheimer's disease through a multivariate multiscale approach. *IEEE Sensors Journal*, 13(9):3284–3292, 2013. IEEE.

251. F. Landini, F. Malerba, and R. Mavilia. The structure and dynamics of networks of scientific collaborations in Northern Africa. *Scientometrics*, 105(3):1787–1807, 2015. Elsevier.

252. J. Lanjouw and M. Schankerman. Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495):441–465, 2004. Wiley Online Library.

253. V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001. APS.

254. Q.V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In *Proc. of the International Conference on Machine Learning (ICML'14)*, pages 1188–1196, Beijing, China, 2014. JMLR.org.

255. J. Lee, S. Yu, K. Park, Y. Park, and Y. Park. Secure three-factor authentication protocol for multi-gateway iot environments. *Sensors*, 19(10):2358, 2019. Multidisciplinary Digital Publishing Institute.

256. M.L. Lee, L.H. Yang, W. Hsu, and X. Yang. XClust: clustering XML schemas for effective integration. In *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM 2002)*, pages 292–299, McLean, Virginia, USA, 2002. ACM Press.

257. P. Lee, H. Su, and F. Wu. Quantitative mapping of patented technology – The case of electrical conducting polymer nanocomposite. *Technological Forecasting and Social Change*, 77(3):466–478, 2010. Elsevier.

258. E.A. Leicht, P. Holme, and M. E. J. Newman. Vertex similarity in networks. *Physical Review Part E*, 73(2):026120, 2006.

259. S.J.J. Leistedt, N. Coumans, M. Dumont, J.P. Lanquart, C.J. Stam, and P. Linkowski. Altered sleep brain functional connectivity in acutely depressed patients. *Human Brain Mapping*, 30(7):2207–2219, 2009. Wiley Online Library.

260. D.A. Levinthal and J.G. March. The myopia of learning. *Strategic Management Journal*, 14(S2):95–112, 1993. Wiley Online Library.

261. Patel N. Culler D. Shenker S. Levis, P. A self-regulating algorithm for code propagation and maintenance in wireless sensor networks. *Symposium on Networked Systems Design and Implementation*.

262. S.R. Levy, K.H. Chiappa, C.J. Burke, and R.R. Young. Early evolution and incidence of electroencephalographic abnormalities in Creutzfeldt-Jakob disease. *Journal of Clinical Neurophysiology*, 3(1):1–21, 1986. LWW.

263. G. Li. Group-based intrusion detection system in wireless sensor networks. *Computer Communications*.

264. G. Li, B.C. Ooi, J. Feng, J. Wang, and L. Zhou. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *Proc. of the International Conference on Management of Data (SIGMOD/PODS'08)*, pages 903–914, Vancouver, Canada, 2008. ACM.

265. J. Li, H. Wang, and S.U. Khan. A semantics-based approach to large-scale mobile social networking. *Mobile Networks and Applications*, 17(2):192–205, 2012. Springer.

266. M. Li, H. Chen, X. Huang, and L. Cui. EasiCrawl: A Sleep-Aware Schedule Method for Crawling IoT Sensors. In *Proc. of the International Conference on Parallel and Distributed Systems (ICPADS'2015)*, pages 148–155, Melbourne, Australia, 2015. IEEE.

267. X. Li, G. Ouyang, and A.R. Douglas. Predictability analysis of absence seizures with permutation entropy. *Epilepsy Research*, 77:70–74, 2007.

268. X. Li, Y. Tian, F. Smarandache, and R. Alex. An extension collaborative innovation model in the context of big data. *International Journal of Information Technology & Decision Making*, 14(01):69–91, 2015. World Scientific.

269. X. Li, Y. Wang, F. Shi, and W. Jia. Crawler for Nodes in the Internet of Things. *ZTE Communications*, 3:009, 2015.

270. C. Lin, G. Li, Z. Shan, and Y. Shi. Thinking and Modeling for Big Data from the Perspective of the I Ching. *International Journal of Information Technology & Decision Making*, 16(06):1451–1463, 2017. World Scientific.

271. C. Lin, J. Wang, and C. Rong. Towards heterogeneous keyword search. In *Proc. of the ACM Turing 50th Celebration Conference-China (ACM TUR-C'17)*, page 46, Shanghai, China, 2017. ACM.

272. F. Lissoni. Academic patenting in Europe: An overview of recent research and new perspectives. *World Patent Information*, 34(3):197–205, 2012. Elsevier.

273. F. Lissoni and E. Miguelez. Patents, Innovation and Economic Geography. Technical report, Groupe de Recherche en Economie Théorique et Appliquée, 2014.

274. J. Liu, X. Zhang, and L. Zhang. Tree pattern matching in heterogeneous fuzzy XML databases. *Knowledge-Based Systems*, 122:119–130, 2017. Elsevier.

275. L. Liu, H. Ma, D. Tao, and D. Zhang. A hierarchical cooperation model for sensor networks supported cooperative work. In *Proc. of the International Conference on Computer Supported Cooperative Work in Design(CSCWD'06)*, pages 1–6, Nanjing, China, 2006. IEEE.

276. P. Liu and H. Xia. Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics*, 103 (1):101–134, 2015.

277. X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Information processing & management*, 41 (6):1462–1480, 2005.

278. Z. Liu, P. Sun, and Y. Chen. Structured search result differentiation. *Proceedings of the VLDB Endowment*, 2(1):313–324, 2009. VLDB Endowment.

279. P. Lo Giudice, N. Mammone, F.C. Morabito, D. Strati, and D. Ursino. A complex network-based approach to detecting and characterizing ictal states in patients with Childhood Absence Epilepsy. In *Proc. of the International Forum on Research and Technologies for Society and Industry (RTSI 2017)*, pages 392–397, Modena, Italy, 2017. IEEE Computer Society.

280. P. Lo Giudice, N. Mammone, F.C. Morabito, D. Ursino, U. Aguglia, V. Cianci, E. Ferlazzo, and S. Gasparini. Usage of network analysis to investigate Periodic Sharp Wave Complexes in EEGs of patients with sporadic CJD. In *Atti del Venticinquesimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati (SEBD'17)*, pages 1–8, Squillace (CZ), Italy, 2017.

281. P. Lo Giudice, P. Russo, and D. Ursino. A new Social Network Analysis-based approach to extracting knowledge patterns about research activities and hubs in a set of countries. *International Journal of Business Innovation and Research*, 17(2):147–186, 2018. Inderscience.

282. S. Lodhi, M. Mirza, and A. Khawaja. Prion Disease: A Review of Novel Transmission Methods and Efforts at Discovering Preventive and Therapeutic Modalities. *Infectious Diseases in Clinical Practice*, 26(1):3–10, 2018. LWW.

283. L. Lovász. Random walks on graphs: A survey. In *Combinatorics, Paul Erdos is Eighty*, pages 1–46. 1993. Springer.

284. L. Lu, D. Chen, X. Ren, Q. Zhang, Y. Zhang, and T. Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016. Elsevier.

285. L.M. Lubango. The effect of co-inventors' reputation and network ties on the diffusion of scientific and technical knowledge from academia to industry in South Africa. *World Patent Information*, 43:5–11, 2015. Elsevier.

286. H. Ma and W. Liu. Progressive Search Paradigm for Internet of Things. *IEEE MultiMedia*, pages 76–86, 2018. IEEE.

287. C. Madera and A. Laurent. The next information architecture evolution: the data lake wave. In *Proc. of the International Conference on Management of Digital EcoSystems (MEDES'16)*, pages 174–180, Hendaye, France, 2016. ACM.

288. J. Madhavan, P.A. Bernstein, and E. Rahm. Generic schema matching with Cupid. In *Proc. of the International Conference on Very Large Data Bases (VLDB 2001)*, pages 49–58, Rome, Italy, 2001. Morgan Kaufmann.

289. T. Maekawa, Y. Yanagisawa, Y. Sakurai, Y. Kishino, K. Kamei, and T. Okadome. Context-aware web search in ubiquitous sensor environments. *ACM Transactions on Internet Technology (TOIT)*, 11(3):12, 2012. ACM.

290. S. Mallat, E. Hkiri, M. Maraoui, and M. Zrigui. Semantic Network Formalism for Knowledge Representation: Towards Consideration of Contextual Information. *International Journal on Semantic Web and Information Systems (IJSWIS'15)*, 11(4):64–85, 2015. IGI Global.

291. B. Malysiak-Mrozek, M. Stabla, and D. Mrozek. Soft and Declarative Fishing of Information in Big Data Lake. *IEEE Transactions on Fuzzy Systems*, 26(5):2732–2747, 2018. IEEE.

292. N. Mammone, L. Bonanno, S. De Salvo, S. Marino, P. Bramanti, A. Bramanti, and F.C. Morabito. Permutation disalignment index as an indirect, EEG-based, measure of brain connectivity in MCI and AD patients. *International journal of neural systems*, 27(05):1750020, 2017. World Scientific.

293. N. Mammone, J. Duun-Henriksen, T.W. Kjaer, M. Campolo, F. La Foresta, and F.C. Morabito. Quantifying the Complexity of Epileptic EEG. In *Smart Innovation, Systems and Technologies: Advances in Neural Networks*, pages 223–232. 2016. Springer.

294. N. Mammone, F. La Foresta, G. Inuso, F.C. Morabito, U. Aguglia, and V. Cianci. Algorithms and topographic mapping for epileptic seizures recognition and prediction. *Frontiers in Artificial Intelligence and Applications*, (204):261–270, 2009. IOS Press.

295. N. Mammone, F. La Foresta, and F.C. Morabito. Discovering network phenomena in the epileptic electroencephalography through permutation entropy mapping. *Frontiers in Artificial Intelligence and Applications*, Neural Nets WIRN10:260–269, 2011. IOS Press.

296. N. Mammone, J.D. Henriksen, T.W. Kjaer, and F.C. Morabito. Differentiating interictal and ictal states in childhood absence epilepsy through Permutation Renyi Entropy. *Entropy*, 17(7):4627–4643, 2015.

297. N. Mammone and F.C. Morabito. Analysis of absence seizure eeg via permutation entropy spatio-temporal clustering. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2011)*, pages 1417–1422, San Jose, CA, USA, 2011. IEEE.

298. N. Mammone, J.C. Principe, F.C. Morabito, D.S. Shiau, and J.C. Sackellares. Visualization and modelling of STLmax topographic brain activity maps. *Journal of neuroscience methods*, 189(2):281–294, 2010. Elsevier.

299. C.D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. 2008. Cambridge University Press Cambridge.

300. J.G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991. INFORMS.

301. G. Marra, F. Ricca, G. Terracina, and D. Ursino. Investigating Information Diffusion in a Multi-Social-Network Scenario via Answer Set Programming. In *Proc. of the*

International Conference On Web Reasoning And Rule Systems (RR 2014), pages 191–196, Athens, Greece, 2014. Lecture Notes in Computer Science. Springer.

302. J. Martinez-Gil and J.F. Aldana-Montes. Semantic similarity measurement using historical google search patterns. *Information Systems Frontiers*, 15(3):399–410, 2013. Springer.
303. S. Maslov and S. Redner. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *Journal of Neuroscience*, 28(44):11103–11105, 2008. Soc Neuroscience.
304. C. McGrath and D. Krackhardt. Network conditions for organizational change. *The Journal of Applied Behavioral Science*, 39(3):324–336, 2003. Sage Publications.
305. M. McPherson, L. Smith-Lovin, and J.M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. JSTOR.
306. O. Mehdi, H. Ibrahim, and L. Affendey. An approach for instance based schema matching with Google similarity and regular expression. *International Arab Journal of Information Technology*, 14(5):755–763, 2017.
307. M. Meyer. What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49(1):93–123, 2000. Springer.
308. E. Miguélez and R. Moreno. Research networks and inventors' mobility as drivers of innovation: evidence from Europe. *Regional Studies*, 47(10):1668–1685, 2013. Taylor & Francis.
309. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
310. T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of the International Conference on Advances in Neural Information Processing Systems (NIPS'13)*, pages 3111–3119, Lake Tahoe, NV, USA, 2013.
311. A.G. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
312. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. American Association for the Advancement of Science.
313. N. Miloslavskaya and A. Tolstoy. Big data, fast data and data lake concepts. *Procedia Computer Science*, 88:300–305, 2016. Elsevier.
314. D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac. Internet of things: Vision, applications and research challenges. *Ad Hoc Networks*, 10(7):1497–1516, 2012. Elsevier.
315. F. Miraglia, F. Vecchio, and P. Rossini. Searching for signs of aging and dementia in EEG through network analysis. *Behavioural Brain Research*, 317:292–300, 2017. Elsevier.
316. S. Misra, R. Barthwal, and M. S. Obaidat. Community detection in an integrated Internet of Things and social network architecture. In *Proc. of IEEE Global Communications Conference (GLOBECOM 2012)*, pages 1647–1652, Anaheim, CA, USA, 2012. IEEE.

317. S. Misra, R. Barthwal, and M.S. Obaidat. Community detection in an integrated Internet of Things and social network architecture. In *Proc. of the Global Communications Conference (GLOBECOM'12)*, pages 1647–1652, Anaheim, CA, USA, 2012. IEEE.
318. F. Montobbio and V. Sterzi. Inventing together: exploring the nature of the Internationalknowledge spillovers in Latin America. *Journal of Evolutionary Economics*, 21(1):53–89, 2011. Springer.
319. F.C. Morabito, M. Campolo, D. Labate, G. Morabito, L. Bonanno, A. Bramanti, S. De Salvo, A. Marra, and P. Bramanti. A longitudinal EEG study of Alzheimer's disease progression based on a complex network approach. *International journal of neural systems*, 25(02):1550005, 2015. World Scientific.
320. F.C. Morabito, M. Campolo, N. Mammone, M. Versaci, S. Franceschetti, F. Tagliavini, V. Sofia, D. Fatuzzo, A. Gambardella, A. Labate, L. Mumoli, G.G. Tripodi, S. Gasparini, V. Cianci, C. Sueri, E. Ferlazzo, and U. Aguglia. Deep Learning representation from electroencephalography of early-stage Creutzfeldt-Jakob Disease and Features for Differentiation from Rapidly Progressive Dementia. *International Journal of Neural Systems*, 3:1650039, 2016. World Scientific.
321. F.C. Morabito, D. Labate, A. Bramanti, F. La Foresta, G. Morabito, I. Palamara, and H. Szu. Enhanced compressibility of EEG signal in Alzheimer's disease patients. *IEEE Sensors Journal*, 13(9):3255–3262, 2013. IEEE.
322. F.C. Morabito, D. Labate, G. Morabito, I. Palamara, and H. Szu. Monitoring and diagnosis of Alzheimer's disease using noninvasive compressive sensing EEG. In *SPIE Defense, Security, and Sensing*, pages 87500Y–87500Y. 2013. International Society for Optics and Photonics.
323. D.V. Moretti. Association of EEG, MRI, and regional blood flow biomarkers is predictive of prodromal Alzheimers disease. *Neuropsychiatric disease and treatment*, 11:2779, 2015. Dove Press.
324. S. Murugesh and A. Jaya. Representing Natural Language Sentences in RDF Graphs to Derive Knowledge Patterns. In *Proc. of the International Conference on Data Engineering and Communication Technology (ICDECT'17)*, pages 701–707, Maharashtra, India, 2017. Springer.
325. A. Nandi and P.A. Bernstein. HAMSTER: Using Search Clicklogs for Schema and Taxonomy Matching. *Proceedings of the VLDB Endowment*, 2(1):181–192, 2009.
326. R. Navigli and S.P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. Elsevier.
327. M.Y. Neufeld and A.D. Korczyn. Topographic distribution of the periodic discharges in Creutzfeldt-Jakob disease (CJD). *Brain Topography*, 4(3):201–206, 1992. Springer.
328. M. Newman and E.A. Leicht. Mixture models and exploratory analysis in networks. *Proc. of the National Academy of Sciences of the United States of America*, 104:9564–9, 2007.
329. M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002. APS.

330. M.E.J. Newman. The structure and function of complex networks. *SIAM review*, pages 167–256, 2003. JSTOR.

331. Liu J. Lyu M. R. Ngai, E. C. H. An efficient intruder detection algorithm against sinkhole attacks in wireless sensor networks. *Computer Communication*.

332. Q.V.H. Nguyen, T.T. Nguyen, V.T. Chau, T.K. Wijaya, Z. Miklos, K. Aberer, A. Gal, and M. Weidlich. SMART: A tool for analyzing and reconciling schema matching networks. In *Proc. of the International Conference on Data Engineering (ICDE'15)*, pages 1488–1491, Seoul, Korea, 2015. IEEE.

333. H. Ning and Z. Wang. Future internet of things architecture: like mankind neural system or social organization framework? *IEEE Communications Letters*, 15(4):461–463, 2011. IEEE.

334. M. Nitti, R. Girau, and L. Atzori. Trustworthiness management in the social internet of things. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1253–1266, 2014. IEEE.

335. A. Nocera and D. Ursino. PHIS: a system for scouting potential hubs and for favoring their "growth" in a Social Internetworking Scenario. *Knowledge-Based Systems*, 36:288–299, 2012. Elsevier.

336. S. Oh, Y. Kim, and S. Cho. An interoperable access control framework for diverse iot platforms based on oauth and role. *Sensors*, 19(8):1884, 2019. Multidisciplinary Digital Publishing Institute.

337. V.P. Oikonomou, A.T. Tzallas, and D.I. Fotiadis. A Kalman filter based methodology for EEG spike enhancement. *Computer Methods and Programs in Biomedicine*, 85(2):101–108, 2007. Elsevier.

338. C. Olston and M. Najork. Web crawling. *Foundations and Trends*, 4(3):175–246, 2010. Now Publishers, Inc.

339. J.P. Onnela, J. Saramäki, J. Kertész, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005. APS.

340. T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010. Elsevier.

341. A. Oram. *Managing the Data Lake*. Sebastopol, CA, USA, 2015. O'Reilly.

342. A. Ortiz, D. Hussein, S. Park, S. Han, and N. Crespi. The cluster between internet of things and social networks: Review and research challenges. *IEEE Internet of Things Journal*, 1(3):206–215, 2014. IEEE.

343. G. Ouyang, J. Li, X. Liu, and X. Li. Dynamic characteristics of absence EEG recordings with multiscale permutation entropy analysis. *Epilepsy Research*, 104(3):246–252, 2013.

344. B. Oyelaran-Oyeyinka, G.O.A. Laditan, and A.O. Esubiyi. Industrial innovation in Sub-Saharan Africa: the manufacturing sector in Nigeria. *Research Policy*, 25(7):1081–1096, 1996. Elsevier.

345. L. Palopoli, L. Pontieri, G. Terracina, and D. Ursino. Intensional and extensional integration and abstraction of heterogeneous databases. *Data & Knowledge Engineering*, 35(3):201–237, 2000.

346. L. Palopoli, D. Rosaci, G. Terracina, and D. Ursino. A graph-based approach for extracting terminological properties from information sources with heterogeneous formats. *Knowledge and Information Systems*, 8(4):462–497, 2005.

347. L. Palopoli, D. Saccà, G. Terracina, and D. Ursino. Uniform techniques for deriving similarities of objects and subschemes in heterogeneous databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):271–294, 2003.

348. L. Palopoli, G. Terracina, and D. Ursino. DIKE: a system supporting the semi-automatic construction of Cooperative Information Systems from heterogeneous databases. *Software Practice & Experience*, 33(9):847–884, 2003.

349. L. Palopoli, G. Terracina, and D. Ursino. Experiences using DIKE, a system for supporting cooperative information system and data warehouse design. *Information Systems*, 28(7):835–865, 2003.

350. Katsaros D. Manolopoulos Y. Papadimitriou, A. Social network analysis and its applications in wireless sensor and vehicular networks. *Conference on e-Democracy*.

351. Y. Papakonstantinou, H. Garcia-Molina, and J. Widom. Object exchange across heterogeneous information sources. In *Proc. of the International Conference on Data Engineering (ICDE'95)*, pages 251–260, Taipei, Taiwan, 1995. IEEE Computer Society.

352. J.-R. Park. Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3-4):213–228, 2009.

353. J.-R. Park and Y. Tosaka. Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly*, 48(8):696–715, 2010.

354. M. Park, H. Oh, and K. Lee. Security risk measurement for information leakage in iot-based smart homes from a situational awareness perspective. *Sensors*, 19(9):2148, 2019. Multidisciplinary Digital Publishing Institute.

355. E. Parvinnia, M. Sabeti, M. Jahromi, and R. Boostani. Classification of EEG Signals using adaptive weighted distance nearest neighbor algorithm. *Journal of King Saud University-Computer and Information Sciences*, 26(1):1–6, 2014. Elsevier.

356. K. Passi, L. Lane, S.K. Madria, B.C. Sakamuri, M.K. Mohania, and S.S. Bhowmick. A model for XML Schema integration. In *Proc. of the International Conference on E-Commerce and Web Technologies (EC-Web 2002)*, pages 193–202, Aix-en-Provence, France, 2002. Lecture Notes in Computer Science, Springer.

357. A. Patel and T.A. Champaneria. Fuzzy logic based algorithm for Context Awareness in IoT for Smart home environment. In *Proc. of the International Conference on Region 10 Conference (TENCON '16)*, pages 1057–1060, Singapore, 2016. IEEE.

358. M. Patella and P. Ciaccia. Approximate similarity search: A multi-faceted problem. *Journal of Discrete Algorithms*, 7(1):36–48, 2009. Elsevier.

359. M. Pavlov and R. Ichise. Finding Experts by Link Prediction in Co-authorship Networks. In *Proc. of the International Workshop on Finding Experts on the Web with Semantics (FEWS 2007)*, pages 42–55, Busan, Korea, 2007.

360. L. Peckeu, N. Delasnerie-Lauprètre, J.P. Brandel, D. Salomon, V. Sazdovitch, J.L. Laplanche, C. Duyckaerts, D. Seilhean, S. Haïk, and J.J. Hauw. Accuracy of diagnosis

criteria in patients with suspected diagnosis of sporadic Creutzfeldt-Jakob disease and detection of 14-3-3 protein, France, 1992 to 2009. *Eurosurveillance*, 22(41), 2017. European Centre for Disease Prevention and Control.

361. C. Perera, Y. Qin, J. Estrella, S. Reiff-Marganiec, and A. Vasilakos. Fog computing for sustainable smart cities: A survey. *ACM Computing Surveys (CSUR)*, 50(3):32, 2017. ACM.
362. C. Perera and A. Vasilakos. A knowledge-based resource discovery for Internet of Things. *Knowledge-Based Systems*, 109:122–136, 2016. Elsevier.
363. C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Context aware computing for the Internet of Things: A survey. *IEEE Communications Surveys & Tutorials*, 16(1):414–454, 2014. IEEE.
364. R. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004. Wiley Online Library.
365. P. Pierleoni, L. Pernini, A. Belli, and L. Palma. An android-based heart monitoring system for the elderly and for patients with heart disease. *International journal of telemedicine and applications*, 2014. Elsevier.
366. G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information processing & management*, 12(5):297–312, 1976. Elsevier.
367. M. Plantié and M. Crampes. Survey on social community detection. In *Social media retrieval*, pages 65–85. 2013. Springer.
368. S.S. Poil, W. De Haan, W.M. van der Flier, H.D. Mansvelder, P. Scheltens, and K. Linkenkaer-Hansen. Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage. *Frontiers in aging neuroscience*, 5:58, 2013. Frontiers.
369. S. Ponten, P. Tewarie, A. Slooter, C. Stam, and E. van Dellen. Neural network modeling of EEG patterns in encephalopathy. *Journal of Clinical Neurophysiology*, 30(5):545–552, 2013. LWW.
370. S.C. Ponten, L. Douw, F. Bartolomei, J.C. Reijneveld, and C.J. Stam. Indications for network regularization during absence seizures: weighted and unweighted graph theoretical analyses. *Experimental Neurology*, 217(1):197–204, 2009. Elsevier.
371. D.V. Prasad, S. Madhusudanan, and S. Jaganathan. uCLUST - A new algorithm for clustering unstructured data. *ARPN Journal of Engineering and Applied Sciences*, 10(5):2108–2117, 2015.
372. Y. Qin, Q. Sheng, N. Falkner, S. Dustdar, H. Wang, and A. Vasilakos. When things matter: A survey on data-centric internet of things. *Journal of Network and Computer Applications*, 64:137–153, 2016. Elsevier.
373. E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.
374. J. Ramirez, J. Gorriz, D. Salas-Gonzalez, A. Romero, M. Lopez, I. Alvarez, and M. Gomez-Rio. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences*, 237:59–72, 2013. Elsevier.

375. S. Rangarajan, H. Liu, H. Wang, and C.L. Wang. Scalable Architecture for Personalized Healthcare Service Recommendation Using Big Data Lake. In *Service Research and Innovation*, pages 65–79. 2015. Springer.

376. A.H. Rasti, M. Torkjazi, R. Rejaie, and D. Stutzbach. Evaluating Sampling Techniques for Large Dynamic Graphs. *Univ. Oregon, Tech. Rep. CIS-TR-08-01*, 2008.

377. D.K. Ravish, S.S. Devi, and S. G. S.G. Krishnamoorthy. Wavelet analysis of EEG for seizure detection: Coherence and phase synchrony estimation. *Biomedical Research*, 2015. Biomedical Research.

378. C. Razafimandimby, V. Loscri, and A.M. Vegni. A neural network and iot based scheme for performance assessment in internet of robotic things. In *Proc. of the International Conference on Internet-of-Things Design and Implementation (IoTDI'16)*, pages 241–246, Orlando, USA, 2016. IEEE.

379. P.J.A. Robson, H.M. Haugh, and B.A.Obeng. Entrepreneurship and innovation in Ghana: enterprising Africa. *Small Business Economics*, 32(3):331–350, 2009. Springer.

380. A.C. Rodrigues, B.S. Machado, L.O.S.F. Caboclo, A. Fujita, L.A. Baccaia, and K. Sameshima. Source and sink nodes in absence seizures. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'16)*, pages 2814–2817, Orlando, FL, USA, 2016. IEEE Society.

381. A. Rodriguez, A. Tosyali, B. Kim, J. Choi, J. Lee, B. Coh, and M. Jeong. Patent Clustering and Outlier Ranking Methodologies for Attributed Patent Citation Networks for Technology Opportunity Discovery. *IEEE Transactions on Engineering Management*, 63(4):426–437, 2016. IEEE.

382. G.A. Ronda-Pupo and L.A. Guerras-Martin. Collaboration network of knowledge creation and dissemination on Management research: ranking the leading institutions. *Scientometrics*, 107(3):917–939, 2016. Springer.

383. G. Rooks, L. Oerlemans, A. Buys, and T. Pretorius. Industrial innovation in South Africa: A comparative study. *South African Journal of Science*, 101(3-4):149–150, 2005. Open Journals Publishing.

384. F. Rotondi, S. Franceschetti, G. Avanzini, and F. Panzica. Altered EEG resting-state effective connectivity in drug-naïve childhood absence epilepsy. *Clinical Neurophysiology*, 127(2):1130–1137, 2016.

385. M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010. Elsevier.

386. G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966. Springer.

387. V. Saccà, M. Campolo, D. Mirarchi, A. Gambardella, P. Veltri, and F. Morabito. On the Classification of EEG Signal by Using an SVM Based Algorithm. In *Multidisciplinary Approaches to Neural Computing*, pages 271–278. 2018. Springer.

388. F. Sadeghi, S.K. Kumar Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proc. of the International Conference on Computer Vision and Pattern Recognition (CVPM'15)*, pages 1456–1464, Boston, MA, USA, 2015.

389. P. K. Saha, U. Maulik, and S. Basu. Advanced Computational Approaches to Biomedical Engineering. *Springer Science Business Media*, 2014. Springer.

390. M. Sahlgren and R. Cöster. Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. In *Proc. of the International Conference on Computational Linguistics (COLING'04)*, page 487, Geneva, Switzerland, 2004.

391. M. Sahlgren and J. Karlgren. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341, 2005.

392. V. Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in biology and medicine*, 41(12):1110–1117, 2011. Elsevier.

393. A. Salazar and L. Zhao. Rhythmic Pattern Extraction by Community Detection in Complex Networks. In *Proc. of the International Conference on Intelligent Systems (BRACIS'14)*, pages 396–401, Sao Paulo, Brazil, 2014. IEEE.

394. Y. Saleem, N. Crespi, M.H. Rehmani, R. Copeland, D. Hussein, and E. Bertin. Exploitation of social IoT for recommendation services. In *Proc. of the World Forum on Internet of Things (WF-IoT'16)*, pages 359–364, Reston, VA, USA, 2016. IEEE.

395. Z. Sankari, H. Adeli, and A. Adeli. Intrahemispheric, interhemispheric, and distal EEG coherence in Alzheimer?s disease. *Clinical Neurophysiology*, 122(5):897–906, 2011. Elsevier.

396. A.F. Santamaria, A. Serianni, P. Raimondo, F. De Rango, and M. Froio. Smart wearable device for health monitoring in the internet of things (IoT) domain. In *Proc. of the International Conference on Proceedings of the summer computer simulation conference(SCSC'16)*, page 36, Montreal, Canada, 2016. Society for Computer Simulation International.

397. J. Santos, T. Wauters, B. Volckaert, and F. De Turck. Resource provisioning in fog computing: From theory to practice. *Sensors*, 19(10):2238, 2019. Multidisciplinary Digital Publishing Institute.

398. S.F. Sayeedunnissa, A.R. Hussain, and M.A. Hameed. Supervised Opinion Mining of Social Network Data Using a Bag-of-Words Approach on the Cloud. In *Proc. of the International Conference on Bio-Inspired Computing: Theories and Applications (BICTA'12)*, pages 299–309, Gwalior, India, 2012.

399. S.E. Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007.

400. S. Schiaffino and A. Amandi. Intelligent user profiling. In *Artificial Intelligence An International Perspective*, pages 193–216. 2009. Springer.

401. A. Seidel. Citation system for patent office. *Journal of the Patent Office Society*, 31(554):26–31, 1949.

402. Z. Shang, Y. Liu, G. Li, and J. Feng. K-join: Knowledge-aware similarity join. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3293–3308, 2016. IEEE.

403. A. Shemshadi, Q. Sheng, and Y. Qin. Thingseek: A crawler and search engine for the internet of things. In *Proc. of the International Conference on Research and Development in Information Retrieval (SIGIR'2016)*, pages 1149–1152, Pisa, Italy, 2016. ACM.

404. A. Shemshadi, L. Yao, Y. Qin, Q. Sheng, and Y. Zhang. Ecs: A framework for diversified and relevant search in the internet of things. In *Proc. of the International Conference on Web Information Systems Engineering (WISE'2015)*, pages 448–462, Miami, Florida, USA, 2015. Springer.

405. L. Shen, B. Xiong, and J. Hu. Research status, hotspots and trends for information behavior in China using bibliometric and co-word analysis. *Journal of Documentation*, 73(4):618–633, 2017. Emerald Publishing Limited.

406. J. Shin, B. Yim, S. Oh, N. Kim, S. kun Lee, and O. Kim. Redefining periodic patterns on electroencephalograms of patients with sporadic Creutzfeldt–Jakob disease. *Clinical Neurophysiology*, 128(5):756–762, 2017. Elsevier.

407. A. Silva and C. Antunes. Constrained pattern mining in the new era. *Knowledge and Information Systems*, 47(3):489–516, 2016. Springer.

408. Ö. Şimşek and D. Jensen. Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, 105(35):12758–12762, 2008.

409. J. Singh. Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, 51(5):756–770, 2005. INFORMS.

410. J. Singh. Distributed R&D, cross-regional knowledge integration and quality of innovative output. *Research Policy*, 37(1):77–96, 2008. Elsevier.

411. K. Singh, K. Paneri, A. Pandey, G. Gupta, G. Sharma, P. Agarwal, and G. Shroff. Visual bayesian fusion to navigate a data lake. In *Proc. of the International Conference on Information Fusion (FUSION'16)*, pages 987–994, 2016. IEEE.

412. K. Singh and V. Singh. Answering graph pattern query using incremental views. In *Proc. of the International Conference on Computing (ICCCA'16)*, pages 54–59, Greater Noida, India, 2016. IEEE.

413. J.V. Sobral, J.J. Rodrigues, R.A. Rabêlo, J. Al-Muhtadi, and V. Korotaev. Routing protocols for low power and lossy networks in internet of things applications. *Sensors*, 19(9):2144, 2019. Multidisciplinary Digital Publishing Institute.

414. O. Sporns and Rolf R. Kötter. Motifs in brain networks. *PLoS Computational Biology*, 2(11):e369, 2004. Public Library of Science.

415. O. Sporns and J.D. Zwi. The small world of the cerebral cortex. *Neuroinformatics*, 2(2):145–162, 2004. Springer.

416. C.J. Stam, W. De Haan, A. Daffertshofer, B. Jones, I. Manshanden, A. Van Walsum, T. Montez, J. Verbunt, J. De Munck, B. Van Dijk, et al. Graph theoretical analysis of magnetoencephalographic functional connectivity in Alzheimer's disease. *Brain*, 132(1):213–224, 2009. Oxford Univ Press.

417. C.J. Stam, B. Jones, G. Nolte, M. Breakspear, and P. Scheltens. Small-world networks and functional connectivity in Alzheimer's disease. *Cerebral cortex*, 17(1):92–99, 2007. Oxford Univ Press.

418. C.J. Stam, Y. Van Der Made, Y. Pijnenburg, and P. Scheltens. EEG synchronization in mild cognitive impairment and Alzheimer's disease. *Acta Neurologica Scandinavica*, 108(2):90–96, 2003. Wiley Online Library.

419. C.J. Stam and J.C. Reijneveld. Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomedical Physics*, 1(1):1, 2007. BioMed Central.

420. C.J. Stam, T.C.A.M. Van Woerkom, and R.W.M. Keunen. Non-linear analysis of the electroencephalogram in Creutzfeldt-Jakob disease. *Biological Cybernetics*, 77(4):247–256, 1997. Springer.

421. B.J. Steinhoff, S. Racker, G. Herrendorf, S. Poser, S. Grosche, I. Zerr, H. Kretzschmar, and T. Weber. Accuracy and reliability of periodic sharp wave complexes in Creutzfeldt-Jakob disease. *Archives of Neurology*, 53(2):162–166, 1996. American Medical Association.

422. B.J. Steinhoff, I. Zerr, M. Glatting, W. Schulz-Schaeffer, S. Poser, and H.A. Kretzschmar. Diagnostic value of periodic complexes in Creutzfeldt–Jakob disease. *Annals of Neurology*, 56(5):702–708, 2004. Wiley Online Library.

423. K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37, 1989. Elsevier.

424. C. Sternitzke, A. Bartkowski, and R. Schramm. Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, 30(2):115–131, 2008. Elsevier.

425. I. Stojmenovic and S. Olariu. Data-centric protocols for wireless sensor networks. *Handbook of sensor networks: algorithms and architectures*, pages 417–456, 2005. Wiley.

426. D. Stutzback, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. In *Proc. of the International Conference on Internet Measurements (IMC'2006)*, pages 27–40, Rio De Janeiro, Brasil, 2006. ACM.

427. F. Su, J. Wang, B. Deng, X.L. Wei, Y.Y. Chen, C. Liu, and H.Y. Li. Adaptive control of Parkinson's state based on a nonlinear computational model with unknown parameters. *International Journal of Neural Systems*, 25(01):1450030, 2015. World Scientific.

428. G. Suciu, S. Halunga, A. Vulpe, and V. Suciu. Generic platform for IoT and cloud computing interoperability study. In *Proc. of the International Symposium on Signals, Circuits and Systems (ISSCS'13)*, pages 1–4, Iasi, Romania, 2013. IEEE.

429. D. Sun, G. Zhang, S. Yang, W. Zheng, S.U. Khan, and K. Li. Re-Stream: Real-time and energy-efficient resource scheduling in big data stream computing environments. *Information Sciences*, 319:92–112, 2015.

430. J. Szymanski. Comparative Analysis of Text Representation Methods Using Classification. *Cybernetics and Systems*, 45(2):180–199, 2014.

431. P.P. Talukdar, Z. Ives, and F. Pereira. Automatically incorporating new sources in keyword search-based data integration. In *Proc. of the International Conference on SIGMOD International Conference on Management of data (SIGMOD/PODS'10)*, pages 387–398, Indianapolis, IN, USA, 2010. ACM.

432. A. Tani, L. Candela, and D. Castelli. Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, 49(6):1194–1205, 2013.

433. K. Tei and L. Gurgen. ClouT: Cloud of things for empowering the citizen clout in smart cities. In *Proc. of the World Forum on Internet of Things (WF-IoT'2014)*, pages 369–370, Seoul, South Korea, 2014. IEEE.

434. I.G. Terrizzano, P.M. Schwarz, M. Roth, and J.E. Colino. Data Wrangling: The Challenging Journey from the Wild to the Lake. In *Proc. of the International Conference on Innovative Data Systems Research (CIDR'15)*, Asilomar, CA, USA, 2015.
435. M. Tortoriello and D. Krackhardt. Activating cross-boundary knowledge: The role of Simmelian ties in the generation of innovations. *Academy of Management Journal*, 53(1):167–181, 2010. Academy of Management.
436. N. Tran, Q. Sheng, M. Babar, and L. Yao. Searching the Web of Things: State of the Art, Challenges, and Solutions. *ACM Computing Surveys (CSUR)*, 50(4):55, 2017. ACM.
437. R.D. Traub and T.A. Pedley. Virus-induced electrotonic coupling: Hypothesis on the mechanism of periodic EEG discharges in Creutzfeldt-Jakob disease. *Annals of Neurology*, 10(5):405–410, 1981. Wiley Online Library.
438. C. Tsai, C. Lai, and A. Vasilakos. Future Internet of Things: open issues and challenges. *Wireless Networks*, 20(8):2201–2217, 2014. Springer.
439. M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. Sebastopol, CA, USA, 2011. O'Reilly Media, Inc.
440. M. Tubaishat, J. Yin, B. Panja, and S. Madria. A secure hierarchical model for sensor network. *ACM Sigmod Record*, 33(1):7–13, 2004. ACM.
441. C.J. Van Rijsbergen. *Information Retrieval*. 1979. Butterworth.
442. F. Vecchio, F. Miraglia, C. Marra, D. Quaranta, M. Vita, P. Bramanti, and P. Rossini. Human brain networks in cognitive decline: a graph theoretical analysis of cortical connectivity from EEG data. *Journal of Alzheimer's Disease*, 41(1):113–127, 2014. IOS Press.
443. B. Verspagen. Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, 10(01):93–115, 2007. World Scientific.
444. F. Vialatte, A. Cichocki, G. Dreyfus, T. Musha, S.L. Shishkin, and R. Gervais. Early detection of Alzheimerós disease by blind source separation, time frequency representation, and bump modeling of EEG signals. In *Proc. of the International Conference on Artificial Neural Networks (ICANN'05)*, pages 683–692, Warsaw, Poland, 2005. Lecture Notes in Computer Science, Springer.
445. J.R. Villar, P. Vergara, M. Menéndez, E. de la Cal, V.M. González, and J. Sedano. Generalized Models for the Classification of Abnormal Movements in Daily Life and its Applicability to Epilepsy Convulsion Recognition. *International Journal of Neural Systems*, 26(06):1650037, 2016. World Scientific.
446. C.S. Wagner and L. Leydesdorff. Network structure, self-organization, and the growth of the internationalcollaboration in science. *Research Policy*, 34(10):1608–1618, 2005. Elsevier.
447. C. Walker and H. Alrehamy. Personal data lake with data gravity pull. In *Proc. of the International Conference on Big Data and Cloud Computing (BDCloud'15)*, pages 160–167, Dalian, China, 2015. IEEE.

448. J. Wan, J. Liu, Z. Shao, A. Vasilakos, M. Imran, and K. Zhou. Mobile crowd sensing for traffic prediction in internet of vehicles. *Sensors*, 16(1):88, 2016. Multidisciplinary Digital Publishing Institute.

449. J. Wan, S. Tang, Z. Shu, D. Li, S. Wang, M. Imran, and A. Vasilakos. Software-defined industrial internet of things in the context of industry 4.0. *IEEE Sensors Journal*, 16(20):7373–7380, 2016. IEEE.

450. F. Wang, Z. Wang, Z. Li, and J.R. Wen. Concept-based Short Text Classification and Ranking. In *Proc. of the International Conference on Information and Knowledge Management (CIKM'14)*, pages 1069–1078, Shangai, China, 2014. ACM.

451. H. Wang, Z. Xu, H. Fujita, and S. Liu. Towards felicitous decision making: An overview on challenges and trends of Big Data. *Information Sciences*, 367:747–765, 2016.

452. J. Wang, J. Li, and J.X. Yu. Answering tree pattern queries using views: a revisit. In *Proc. of the International Conference on Extending Database Technology (EDBT/ICDT'11)*, pages 153–164, Uppsala, Sweden, 2011. ACM.

453. L. Wang. Using the relationship of shared neighbors to find hierarchical overlapping communities for effective connectivity in IoT. In *Proc. of the International Conference on Pervasive Computing and Applications (ICPCA'11)*, pages 400–406, Port Elizabeth, South Africa, 2011. IEEE.

454. L. Wang, C. Zhu, Y. He, Y. Zang, Q. Cao, H. Zhang, Q. Zhong, and Y. Wang. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Human Brain Mapping*, 30(2):638–649, 2009. Wiley Online Library.

455. L.H. Wang, R.C. Bucelli, E. Patrick, D. Rajderkar, E. Alvarez III, M.M. Lim, G. DeBruin, V. Sharma, S. Dahiya, R.E. Schmidt an T.S. Benzinger, B.A. Ward, and B.M. Ances. Role of magnetic resonance imaging, cerebrospinal fluid, and electroencephalogram in diagnosis of sporadic Creutzfeldt-Jakob disease. *Journal of Neurology*, 260(2):498–506, 2013. Springer.

456. P.S. Wang, Y.T. Wu, C.I. Hung, S.Y. Kwan, S. Teng, and B.W. Soong. Early detection of periodic sharp wave complexes on EEG by independent component analysis in patients with Creutzfeldt-Jakob disease. *Journal of Clinical Neurophysiology*, 25(1):25–31, 2008. LWW.

457. R. Wang, J. Wang, H. Yu, X. Wei, C. Yang, and B. Deng. Power spectral density and coherence analysis of Alzheimerós EEG. *Cognitive neurodynamics*, 9(3):291–304, 2015. Springer.

458. S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge, UK, 1994. Cambridge University Press.

459. D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. Nature Publishing Group.

460. D. Wei, X. Deng, X. Zhang, Y. Deng, and S. Mahadevan. Identifying influential nodes in weighted networks based on evidence theory. *Physica A: Statistical Mechanics and its Applications*, 392(10):2564–2575, 2013. Elsevier.

461. P.D. Welch. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

462. World Health Organization (WHO). Consensus on criteria for diagnosis of sporadic CJD. *Weekly Epidemiological Record*, 73:361–365, 1998.

463. H.G. Wieser, K. Schindler, and D. Zumsteg. EEG in Creutzfeldt–Jakob disease. *Clinical Neurophysiology*, 117(5):935–951, 2006. Elsevier.

464. H.G. Wieser, U. Schwarz, T. Blättler, C. Bernoulli, M. Sitzler, K. Stoeck, and M. Glatzel. Serial EEG findings in sporadic and iatrogenic Creutzfeldt–Jakob disease. *Clinical Neurophysiology*, 115(11):2467–2478, 2004. Elsevier.

465. G. Wittenbaum, A. Hubbell, and C. Zuckerman. Mutual enhancement: Toward an understanding of the collective preference for shared information. *Journal of personality and social psychology*, 77(5):967, 1999. American Psychological Association.

466. World Intellectual Property Organization. *Patent-Based Technology Analysis Report - Alternative Energy Technology*. Geneva, Switzerland, 2009.

467. K. Xu, Y. Qu, and K. Yang. A tutorial on the internet of things: From a heterogeneous network integration perspective. *IEEE Network*, 30(2):102–108, 2016. IEEE.

468. J.Z. Yan. Abnormal cortical functional connections in Alzheimer's disease: analysis of inter-and intra-hemispheric EEG coherence. *Journal of Zhejiang University Science B*, 6(4):259–264, 2005. Springer.

469. G. Yang, G. Li, C. Li, Y. Zhao, J. Zhang, T. Liu, D. Chen, and M. Huang. Using the comprehensive patent citation network (CPC) to evaluate patent value. *Scientometrics*, 105(3):1319–1346, 2015. Springer.

470. S. Ye, J. Lang, and F. Wu. Crawling online social graphs. In *Proc. of the International Asia-Pacific Web Conference (APWeb'10)*, pages 236–242, Busan, Korea, 2010. IEEE.

471. J. Yu, Z. Tsai. A Framework of Machine Learning Based Intrusion Detection for Wireless Sensor Networks. *IEEE International Conference on Sensor Networks*.

472. M. Yu, A. Gouw, A. Hillebrand, B. Tijms, C. Stam, E. van Straaten, and Y. Pijnenburg. Different functional connectivity and network topology in behavioral variant of frontotemporal dementia and Alzheimer's disease: an EEG study. *Neurobiology of aging*, 42:150–162, 2016. Elsevier.

473. Y. Yuan, G. Wang, L. Chen, and B. Ning. Efficient pattern matching on big uncertain graphs. *Information Sciences*, 339:369–394, 2016. Elsevier.

474. N. Zamzami and A. Schiffauerova. The impact of individual collaborative activities on knowledge creation and transmission. *Scientometrics*, 111(3):1–29, 2017. Springer.

475. G. Zanusso, S. Monaco, M. Pocchiari, and B. Caughey. Advanced tests for early and accurate diagnosis of Creutzfeldt-Jakob disease. *Nature Reviews Neurology*, 12(6):325–333, 2016. Nature Research.

476. H. Zardi, L.B. Romdhane, and Z. Guessoum. Efficiently mining community structures in weighted social networks. *International Journal of Data Mining, Modelling and Management*, 8(1):32–61, 2016. Inderscience Publishers (IEL).

477. J. Zhang, Y. Jiang. Bibliographical review on reconfigurable fault-tolerant control systems. *Annual Reviews in Control*, 81:111–124, 2017. Elsevier.

478. P. Zhang, Y. Liu, F. Wu, and B. Tang. Matching state estimation scheme for content-based sensor search in the Web of things. *International Journal of Distributed Sensor Networks*, 11(11):326780, 2015. SAGE Publications.

479. Y. Zhang, D. Raychadhuri, L. Grieco, E. Baccelli, J. Burke, R. Ravindran, G. Wang, A. Lindgren, B. Ahlgren, and O. Schelen. Requirements and Challenges for IoT over ICN. *https://tools.ietf.org/html/draft-zhang-icnrg-icniot-requirements-00*, 2015. IETF Internet-Draft.

480. Y. Zhang, D. Raychadhuri, R. Ravindran, and G. Wang. ICN based Architecture for IoT. *https://tools.ietf.org/html/draft-zhang-iot-icn-challenges-02*, 2013. IRTF contribution.

481. F. Zhao, Z. Sun, and H. Jin. Topic-centric and semantic-aware retrieval system for internet of things. *Information Fusion*, 23:33–42, 2015. Elsevier.

482. X. Zhao, Y. Yu, Z. Zhao, and J. Xu. Comparison Between Sporadic and Misdiagnosed Sporadic Creutzfeldt–Jakob Disease: A Report of Two Cases. *Cell biochemistry and biophysics*, 72(2):311–315, 2015. Springer.

483. J. Zhou, Z. Cao, X. Dong, and A. Vasilakos. Security and privacy for cloud-based IoT: Challenges. *IEEE Communications Magazine*, 55(1):26–33, 2017. IEEE.

484. G. Zhu, Y. Li, P.P. Wen, and S. Wang. Classifying epileptic EEG signals with delay permutation entropy and multi-scale k-means. *Advances in Experimental Medicine and Biology*, 823:143–157, 2015.

485. Y. Zhuang, Y. Wang, J. Shao, L. Chen, W. Lu, J. Sun, B. Wei, and J. Wu. D-Ocean: an unstructured data management system for data ocean environment. *Frontiers of Computer Science*, 10(2):353–369, 2016. Springer.

Ringraziamenti

Ho visto un sufficiente numero di tesi in questi ultimi quattro anni per sapere che spesso si inizia questa particolare sezione con un "è forse la parte più difficile della tesi", vero. Vero perchè i ringraziamenti rappresentano, senza ombra di dubbio, la fine del percorso.

Nel mio caso è - purtroppo - particolarmente vero. Pochi sono i capisaldi che accompagnano la mia vita passata e la mia vita futura. E da qui, seppur in parte, ecco spiegata la dedica iniziale.

Queste ultime pagine di scritto rappresentano la chiusura di un lunghissimo percorso, la conclusione di un periodo meraviglioso della mia vita, quello da "studente universitario". Sono stati anni meravigliosi e, anche vista l'enormità di tutto quello che mi sto lasciando alle spalle, mi sono preso tutto il tempo necessario per essere pronto a chiudere questo capitolo.

Vorrei ringraziare seguendo un po' il flusso temporale di questo dottorato. Alcuni potrebbero esserci più volte, per altri potrebbe essere il contrario. Il risultato finale potrebbe sembrare simile ad un lungo elenco, ma portate pazienza (e poi, come forse mi avreste sentito dire, i ringraziamenti sono un modo per dare indietro un briciolo di quanto si è ricevuto).

Il buon Foscolo diceva "non ringraziate il vostro professore, perchè avrà fatto solo il suo lavoro" e seguirò il suo consiglio. Non ringrazio quindi il mio tutor, entrambi sappiamo quali siano state le sfide e gli ostacoli superati durante questi 4 anni e avrebbe poco senso spendere altre parole.

Ringrazio invece - e senza alcuna esitazione - Mimmo. Aver lavorato insieme a te è stato senza alcun dubbio un privilegio ed un piacere enorme. Sia umanamente che professionalmente sei stato per me una guida ed essere così diversi e così aperti al confronto è stato per me motivo di grande orgoglio. Porto con me tantissimo di quello che mi hai insegnato e, la qualità di questo bagaglio, mi è stata chiarissima fin dal primo momento fuori dall'ambiente universitario.

Ringrazio chi era con me al Ciroma, la sera che ho festeggiato il superamento dell'esame. Sono le stesse persone che mi hanno supportato per gran parte della mia vita, ed a tutti loro devo moltissimo. Grazie a voi sono diventato la persona che sono, voi avete fatto maturare e cementare tutte le mie certezze.

Ringrazio Diego, Giacinto, Carlo, Nico ed Enrico, per tutti i pranzi e le discussioni all'università. Ancora oggi le fazioni sul "reddito di nascita" risultano divise, ma fortunatamente non hanno intaccato il nostro rapporto.

Grazie a Davo, Marco, Rocco, Sciop, Dona, Ale e Luca per i viaggi all'estero, le avventure su motorini improbabili e le sbronze a furia di shottini da 1euro.

Grazie a Lorenzo, Enrico, Peppe e Andrea per tutto il tempo trascorso insieme in laboratorio, per le risate e per le giornate spensierate condivise.

Una nota speciale per Musarellaaaaaa: alla fine, contro ogni ipotesi iniziale, abbiamo condiviso solo un piccolo pezzettino del percorso. Non avrei potuto comunque pensare a nulla di meglio.

Grazie a Davo, Peppe (x2), Rosario, Antonella, Davide, Gianni (x2), Bruno, Pippo e tutti quelli che hanno aiutato me e Lorenzo nella gestione del laboratorio nell'ultimo periodo.

Grazie a Davo, Luca e Cris per tutto il sostegno che mi hanno dato nel momento più brutto che io abbia mai passato. Mi avete letteralmente permesso di respirare quando ne avevo più bisogno. Mi sono reso conto solo di recente di che tipo di sforzo vi abbia richiesto, e non esistono parole che possano esprimere appieno quale sia la mia immensa gratitudine.

Per lo stesso motivo ringrazio anche Donato, Alessandro, Ciccio, Fabio, Resena, Peppe, Giacinto, n'altro Peppe, Lorenzo, Fede, Martina, Giulia, Gigi, Mando, Giovanni e chiunque mi abbia dedicato il suo tempo ed i suoi pensieri.

Ringrazio Diego, il boss, per avermi accolto e ospitato quasi facendo finta che glielo avessi chiesto. Per avermi pazientemente spronato e pungolato su tante tematiche.

E con lui ci sono Carlo, Peppe Cuzzola, Ale, Andrea, Peppe Battista. Il nocciolo duro delle persone che stanno accompagnando il mio tempo milanese.

Ringrazio i colleghi di Deloitte (solo i più stretti, altrimenti non finisco più), per come sono stato accolto e per tutti i momenti passati insieme. Avete il potere di rendere divertente il luogo di lavoro, ed anche di farmi ingozzare di arancine e/o arancini. Quindi grazie infinite a Marmariani, CL7, Luigi, Davide, Vinz, Giuseppe, Diego, Francesca, Alessandro, Marco, Olli, Umbe, Alessio, Daniele e Stefano, Matti, Valentina V e Valentina C.

Grazie a Daniel, Richy, Fra, Vane, Davide, Genti, Oscar, Martina, Mariele e Claudia per tutti gli allenamenti, il divertimento, i pugni e soprattutto le birre.

Grazie a Peppe ed Andrea, per ogni pietanza meravigliosa che vi ho visto mangiare e per il (poco) tempo trascorso in casa insieme.

Grazie a Cris, per tutti i sushi e le lunghe conversazioni domenicali. Grazie a Diego, per tutti i manicaretti che mi ha preparato. Entrambi questi grazie sono riduttivi, non descrivono adeguatamente la qualità del tempo passato insieme, ma penso che voi sappiate quanto vale per me questo tempo.

Grazie alla mia Famiglia. Angela, Stella, Leander e Nico. Grazie per tutti gli sforzi, il tempo ed i pensieri che mi dedicate.

Ringrazio, infine, tutti i colleghi fin qui incontrati e tutte le persone che mi sono state vicine in questi anni; senza di voi oggi non sarei ciò che sono.