



Politecnico
di Bari

Department of Electrical and Information Engineering
INDUSTRY 4.0 Ph.D. Program
SSD: FIS/07–APPLIED PHYSICS

Final Dissertation

From smart firms to smart consumers:
Complex Systems and Machine Learning
for Industry 4.0

by

De Nicolò Francesco:

Supervisor:

Prof. Nicola Amoroso

Coordinator of Ph.D. Program:

Prof. Caterina Ciminelli

Course n°36, 01/11/2020-31/10/2023



LIBERATORIA PER L'ARCHIVIAZIONE DELLA TESI DI DOTTORATO

Al Magnifico Rettore
del Politecnico di Bari

Il sottoscritto DE NICOLA' FRANCESCO nato a PUTIGNANO (BA) il 25/04/1989
residente a PUTIGNANO (BA) in via FRATELLI BANDIERA, 6 e-mail francesco.denicolo@poliba.it
iscritto al 3° anno di Corso di Dottorato di Ricerca in INDUSTRIA 4.0 ciclo 36

ed essendo stato ammesso a sostenere l'esame finale con la prevista discussione della tesi dal titolo:

FROM SMART FIRMS TO SMART CONSUMERS: COMPLEX SYSTEMS AND MACHINE LEARNING FOR INDUSTRY 4.0

DICHIARA

- 1) di essere consapevole che, ai sensi del D.P.R. n. 445 del 28.12.2000, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni;
- 2) di essere iscritto al Corso di Dottorato di ricerca INDUSTRIA 4.0 ciclo 36, corso attivato ai sensi del "Regolamento dei Corsi di Dottorato di ricerca del Politecnico di Bari", emanato con D.R. n.286 del 01.07.2013;
- 3) di essere pienamente a conoscenza delle disposizioni contenute nel predetto Regolamento in merito alla procedura di deposito, pubblicazione e autoarchiviazione della tesi di dottorato nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica;
- 4) di essere consapevole che attraverso l'autoarchiviazione delle tesi nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica del Politecnico di Bari (IRIS-POLIBA), l'Ateneo archiverà e renderà consultabile in rete (nel rispetto della Policy di Ateneo di cui al D.R. 642 del 13.11.2015) il testo completo della tesi di dottorato, fatta salva la possibilità di sottoscrizione di apposite licenze per le relative condizioni di utilizzo (di cui al sito <http://www.creativecommons.it/Licenze>), e fatte salve, altresì, le eventuali esigenze di "embargo", legate a strette considerazioni sulla tutelabilità e sfruttamento industriale/commerciale dei contenuti della tesi, da rappresentarsi mediante compilazione e sottoscrizione del modulo in calce (Richiesta di embargo);
- 5) che la tesi da depositare in IRIS-POLIBA, in formato digitale (PDF/A) sarà del tutto identica a quelle **consegnate**/inviata/da inviarsi ai componenti della commissione per l'esame finale e a qualsiasi altra copia depositata presso gli Uffici del Politecnico di Bari in forma cartacea o digitale, ovvero a quella da discutere in sede di esame finale, a quella da depositare, a cura dell'Ateneo, presso le Biblioteche Nazionali Centrali di Roma e Firenze e presso tutti gli Uffici competenti per legge al momento del deposito stesso, e che di conseguenza va esclusa qualsiasi responsabilità del Politecnico di Bari per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
- 6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto il Politecnico di Bari ed i suoi funzionari sono in ogni caso esenti da responsabilità di qualsivoglia natura: civile, amministrativa e penale e saranno dal sottoscritto tenuti indenni da qualsiasi richiesta o rivendicazione da parte di terzi;
- 7) che il contenuto della tesi non infrange in alcun modo il diritto d'Autore né gli obblighi connessi alla salvaguardia di diritti morali od economici di altri autori o di altri aventi diritto, sia per testi, immagini, foto, tabelle, o altre parti di cui la tesi è composta.

Luogo e data BARI, 27/12/2023

Firma 

Il/La sottoscritto, con l'autoarchiviazione della propria tesi di dottorato nell'Archivio Istituzionale ad accesso aperto del Politecnico di Bari (POLIBA-IRIS), pur mantenendo su di essa tutti i diritti d'autore, morali ed economici, ai sensi della normativa vigente (Legge 633/1941 e ss.mm.ii.),

CONCEDE

- al Politecnico di Bari il permesso di trasferire l'opera su qualsiasi supporto e di convertirla in qualsiasi formato al fine di una corretta conservazione nel tempo. Il Politecnico di Bari garantisce che non verrà effettuata alcuna modifica al contenuto e alla struttura dell'opera.
- al Politecnico di Bari la possibilità di riprodurre l'opera in più di una copia per fini di sicurezza, back-up e conservazione.

Luogo e data BARI, 27/12/2023

Firma 



Politecnico
di Bari

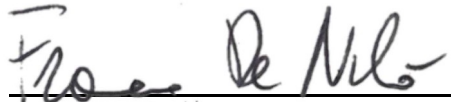
Department of Electrical and Information Engineering
INDUSTRY 4.0 Ph.D. Program
SSD: FIS/07–APPLIED PHYSICS

Final Dissertation

From smart firms to smart consumers:
Complex Systems and Machine Learning
for Industry 4.0

by

De Nicolò Francesco:



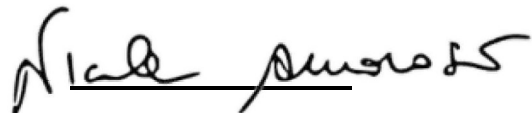
Referees:

Prof. Eleonora Alfinito

Prof. Gastone Castellani

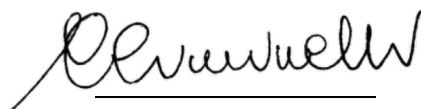
Supervisor:

Prof. Nicola Amoroso



Coordinator of Ph.D Program:

Prof. Caterina Ciminelli



Course n°36, 01/11/2020-31/10/2023

Contents

1	Industry 4.0, Complex Systems and Machine Learning	3
1.1	Complex Systems in Industry 4.0	3
1.2	Taming Complex Systems graph theory	5
1.3	Exploiting Complex Systems Machine Learning models.	6
1.4	Thesis organization...	7
2	How to extract insights for Industry 4.0: graph theory and Machine Learning	9
2.1	Fundamentals of graph theory...	9
2.1.1	Complex Systems and graphs.	9
2.1.2	Graphs' value in modelling Complex Systems.	11
2.1.3	Multi-graph, simple graph and weighted graph	12
2.1.4	Degree and Adjacency matrix	13
2.1.5	Geodetic paths	15
2.1.6	Closeness.	17
2.1.7	Betweenness	18
2.2	Machine Learning algorithms	19
2.2.1	Unsupervised Machine Learning community detection in graphs	19
2.2.2	Supervised Machine Learning algorithms and Explainability	24
3	Startups and consumer reviews: use-cases and why they are important	33
3.1	The startup ecosystem its importance and modelling.	34
3.1.1	The importance of the startup ecosystem for Industry 4.0	34
3.1.2	Countries' innovation ecosystem the StartupBlink ranking	35
3.1.3	Interplay among startups and investors the StartupBlink ranking	38
3.2	Why studying tourists' tastes in Industry 4.0.	39
3.2.1	Tourists' experiences in Apulia from TripAdvisor reviews and rating	40
4	How to boost innovation and customers' satisfaction by employing graph theory and Machine Learning	43
4.1	StartupBlink: equity oriented rethinking through community detection	43
4.1.1	StartupBlink country network	44
4.1.2	Community detection algorithms and Resolution Ratio	44

4.1.3	WDI country communities and StartupBlink rethinking	50
4.2	Crunchbase graph model and forecasting success	55
4.2.1	Modelling the economic interplay	55
4.2.2	Defining and measuring success	57
4.2.3	Strategic elements in the startup ecosystem	58
4.2.4	Forecasting success	61
4.3	TripAdvisor: extracting insights from tourists' reviews	66
4.3.1	From text to number TF-IDF matrix	66
4.3.2	Reviews' classification	67
4.3.3	Strengths and weaknesses of the Apulian tourism offer	68
5	Insights and future perspectives for startups and reviews' analysis in Industry 4.0	71
5.1	Highlighting the best practices in innovation ecosystems through community detection	71
5.2	Graph metrics and startups' success	73
5.3	Unveiling tourists' tastes and needs	74
5.4	Future perspectives of Complex Systems and Machine Learning in Industry 4.0	76
	Appendices	85
	A 2019 StartupBlink ranking	85
	B StartupBlink: community detection and clustering analyses	89
B.1	Spin Glass and Leiden algorithms for feature space exploration	89
B.2	Why not classical clustering methods?	90
B.2.1	Clustering results for StartupBlink countries	93
	C Crunchbase: main features and statistical analyses	97
C.1	Dataset description	97
C.2	Most present elements' attributes	98
C.3	Funding and network metrics: global distribution differences and top fifty ranking	98
C.4	Funding and network metrics: distribution differences for Country, Investor type and Economic category	100
C.5	Statistical analyses	109

Introduction

The Fourth Industrial Revolution also called *Industry 4.0* [1], has re-invented the way firms design, produce and distribute their products. Technologies such as Industrial Internet of Things (IIoT), cloud connectivity and Machine Learning are now deeply intertwined into the production process. This unified and integrated approach to manufacturing results in products, factories, and assets that are connected and intelligent. Accordingly, a firm can be seen as a *Complex System*: every aspect of the firm's activity is strictly linked to each other and their evolution, as well as that of the whole system, depends on these connections. In other words, the evolution of every single productive element of a firm cannot be studied on its own but must be placed in a more holistic framework.

Moreover, there is another often overlooked aspect joining the Fourth Industrial Revolution and Complex Systems. As previously underlined, firms are converting to the Industry 4.0 paradigm with a growing trend of industrial automation that aims at integrating new technologies, creating new business models, increasing productivity and products' quality. Nonetheless, the critical point, before being methodological, is about technology. Startups play a key role in pushing the existing technological limits beyond. As a matter of fact, one technological disruption is the result of many tries and fails, and young firms are more prone to risk and experimentation. The value of Industry 4.0-related startups is also evidenced by their economic value: almost 200 billion euros in the next 2 years [3]. Accordingly, startups represent also one of the main boosts of the economic growth of a country. As a consequence, studies aiming at quantitatively pointing out the most promising startups are gaining more and more ground together with those dealing with the identification of the most strategic elements in setting up an effective economic system supporting innovation. In fact, recently some scholars have introduced the concept of *high-impact entrepreneurship* [4, 5].

In order to accomplish this task, startups must be considered together with their relations with the socio-economic context in which firms rise and grow. This system is called *startup or innovation ecosystem* and presents itself as a *Complex System*. It cannot be studied by classical means but needs suitable mathematical tools.

Accordingly, part of this work is devoted to the study of the startup ecosystem through *graph theory*, the main mathematical instrument used in analyzing *Complex Systems*. This study aims at answering to the following research questions (RQs).

- **RQ-1.** Given the importance of startups in boosting countries' economies, how is the effectiveness of a country's innovation ecosystem influenced by

the socio-economic context in which it grows?

- **RQ-2.** At a greater level of detail, who are the most strategic elements in a startup ecosystem? Is there a relation between the strategic value of a startup in this system and its future success?

The technological revolution characterizing Industry 4.0 has radically changed not only the way firms produce, but even how they engage with consumers. In fact, the latter can establish more interactive relations with firms by posting reviews on online social platforms (TripAdvisor, Amazon, Facebook) through which they express their needs and opinions about products and experiences. These reviews have the possibility to reach and influence the purchasing decisions of other consumers spread over the world. This revolution in consumers' role allows them to be nowadays considered as *co-producers* [6]. Consequently, intercepting and forecasting consumers' needs is actually one of the main keys to firms' success.

Many studies have highlighted that consumers' textual reviews are the best instrument to capture their evaluation of products and experiences [7, 8]: they highlight the products' and services' features customers care about and provide their perceptions in a detailed way through the open-structured form. In fact, in *face-to-face* conversations it is often hard to capture consumers' all evaluation of their experience since they may not articulate feelings, especially in case of negative perception because of worries about breaking the customer-seller relationship [9]. Moreover, the measurement of customers' evaluation through closed-ended survey questions is highly influenced by the way the survey is designed [10, 11].

However, the major challenge in the analysis of written comments is the *information overload* [12]: reading them one by one is time consuming because there is a great number of reviews available online and they contain a substantial number of words.

Accordingly, one fundamental aspect of a firm's activity should be the automated extraction of insights from these reviews. Since this activity deals with textual (i.e. non-structured) data, suitable Machine Learning algorithms and techniques must be employed. In particular, a firm should be able to highlight those aspects that mainly influence a review to be positive or negative. This task can be accomplished using the *explainability tools* of Machine Learning. Explainability represents an active research field, from a theoretical and application point of view [13]. Accordingly, a second part of this work will be devoted to the deployment of suitable Machine Learning tools in order to answer to the following research question:

- **RQ-3.** How can insights from textual data be automatically extracted?

This work is organized as follows. First, it will be underlined the importance of Complex Systems and Machine Learning in the context of the Fourth Industrial Revolution. After showing the fundamental definitions and algorithms of graph theory and Machine Learning, three use-cases, one for each research question, will be first described and then studied through these. Finally, the main findings about these use-cases will be discussed together with future perspectives about the interactions among Industry 4.0, Complex Systems and Machine Learning.

Chapter 1

Industry 4.0, Complex Systems and Machine Learning

This chapter highlights how Complex Systems and Machine Learning can be of invaluable help for firms to take faster and more reliable business decisions. In fact, using these tools they can exploit the great amount of structured (i.e. non-tabular) data coming from different sources like online social networks and consumers' reviews. In particular, in the first section it will be shown how Complex Systems naturally arise in the context of Fourth Industrial Revolution. Then, in the second section, the main theoretical tools for modelling Complex Systems will be introduced. The third section will deal with the main Machine Learning models used to extract insights from both structured and non-structured data. Finally, in the the fourth section the thesis' organization will be shown.

1.1 Complex Systems in Industry 4.0

The Fourth Industrial Revolution (also indicated as *Industry 4.0* [1]) is characterized by an unprecedented technological pervasiveness in every aspect of firms' activities [14]. The most evident result of this revolution is the quantity and variety of data that every productive element of a firm provides, if equipped with appropriate sensors. Such information, when appropriately processed, leads the ability to provide strategic insights about production processes, market trends, and consumers' behaviour, so letting decision-making processes be faster and much more efficient [15, 16].

The epochal significance of Industry 4.0 has been well highlighted by Klaus Schwab (Executive Chairman of the World Economic Forum) in 2017. He identified the reasons why it should be considered responsible for revolutionizing the way we think about productive activities and interpret social relations [17]. In particular, he emphasizes how the role of the consumer is radically changing: the increasing access to knowledge and information makes the relationships between companies and consumers more interactive, making the latter crucial in

the production process, to the extent that they can be considered *co-producers*. For example users' reviews about experiences and products posted in online social platforms can deeply influence the decision-making process of other consumers as widely confirmed by both academic research and practice [18, 20]. In particular, research in tourism, one of the most profitable economic activity, has highlighted online reviews as a major driver of brand choice and sales [21], hotel performance [22], hotel bookings [23] and destination choice [24]. It should be underlined that their effect on guests' satisfaction [25] has opened the discussion on the quantitative analysis of the hospitality experience [26].

Moreover this phenomenon makes the consumer-firm relations much more complicated as each individual consumer or company both influences and is influenced by all the other people and firms it comes into contact with. In other words, users and firms cannot be considered as isolated from the socio-economic context in which they live and since these interactions may massively orient tastes and buying habits, they could determine the success or breakdown of a company [6]. Therefore it is of paramount importance to deeply understand these systems composed of mutually interacting elements (firms and consumers) that influence each other determining the properties and evolution of both the whole system and of the single elements.

These systems are collectively called *Complex Systems*. Complex Systems represent the cornerstone of many branches of scientific research, owing to the generality of their definition: a great number of natural and artificial systems fall within. For example:

- In the biomedical field, neurons, the basis of the nervous system, are cells with mutual connections (synapses) through which their communication occurs (i.e. the passage of electrical impulses). For other examples of complex systems in Biology, refer to the work of Jeong et al. [27] on cell networks and [28] on the network modelling of food chains.
- In the field of computer science and technology, we can consider Internet and the World Wide Web (see, for example, the works of Capocci et al. [29], Vazquez et al. [30] and Pastor-Satorras et al. [31] on the structure of Internet). In particular, Internet is composed of computers and routers that exchange data using electromagnetic signals. Internet is geographically distributed worldwide and is one of the most extensively studied Complex Systems. The World Wide Web on the other hand, is a service that utilizes data transfer provided by Internet and consists of a collection of contents (web pages) linked to each other by specific links (also known as hyperlinks) through which users can navigate from one content to another.
- In the field of social sciences, the most prominent example of a Complex System is provided by Social Networks, i.e., groups of people connected by friendship, kinship, or other types of relationships (see the works of Newman [32, 33] on the structure of Social Networks). The use of online Social Networks (like Facebook, Instagram, Tencent Weibo) has seen significant growth in recent years, making large amounts of data available for analysis.

Considering the ever-increasing quantity and heterogeneity of available data,

the most suitable analytical tools to be used to identify hidden patterns and find relationships among data are the Machine Learning algorithms [34, 35, 36, 37].

On the other hand, as previously underlined, relationships and interactions among people and firms drive buying dynamics and the path to success of firms: the absence of this essential aspect in data representation makes predictive results less reliable [38]. In other words, to enhance the predictive power of forecasting tools on Complex Systems, it is necessary to model the relationships and feed Machine Learning algorithms with such additional information.

According to these aspects of Industry 4.0, the research path followed in this work is composed of two main pillars: Complex Systems modelling for modelling relations in Complex Systems and Machine Learning for forecasting tasks. In the next section the theoretical backgrounds of Complex Systems modelling will be presented, while the relation among Machine Learning and Complex Systems will be deepened in the third section.

1.2 Taming Complex Systems graph theory

Research on Complex Systems has posed new challenges to the conventional methods of problem-solving. As previously stated, a Complex System is composed of interacting elements that influence each other and together determine the properties and evolution both of the whole system and of the single elements. For example, knowing how a neuron works is not sufficient to describe the brain's functions: a model taking care of the interactions between neurons must be used to describe brain properties. Accordingly, in order to determine the properties characterizing a Complex System, it must not be considered as a separate collection of items and a modelling method should be used that considers the properties of the entire system together with those of the individual parts. This means that classical statistical analysis cannot capture all the features of these systems. Accordingly, correctly modelling Complex Systems plays the most important part in their analysis.

It should be underlined that, because of the pervasiveness of Complex Systems in both nature and human-generated systems, they have been objects of investigation of various branches of science. Historically, this differentiation has led to an independent evolution of concepts and methods for such systems in distinct research fields. For example, well-acknowledged tools in the Social Sciences were not known in Physics or Chemistry, and vice versa. This has hindered the development of a unified body of knowledge about Complex Systems, at least until a few decades ago when the Science of Complexity established itself as a separate scientific field. Accordingly, the main mathematical tool used to model Complex Systems is *graph theory*, or *network theory* [39]. The origin of graphs dates back to the pioneering work of Euler (1736) about the *problem of the seven bridges of Königsberg* [32]. A graph (or network), G , is defined as a couple (V, L) , where V is a non-empty set and L is a set of couples of elements of V [40]. The elements of V are called *vertices* (or *nodes*) while the couples in L are called *links* (or *edges*). Graphically, a network may be depicted as a set of points (the nodes) linked by lines representing the edges of the network. Figure 1.1 is an example of a graphical representation of a network.

It is evident from graph's definition how it is well suited to model Complex Systems, since their elements can be considered as the nodes of a graph while the

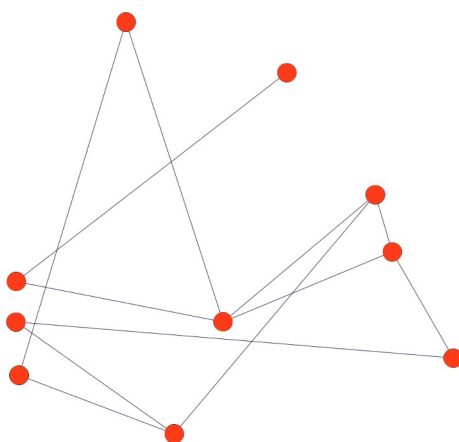


Figure 1.1 Graphical representation of a network

links can model their relationships (friendship, physical or electromagnetic connections) Since networks model interacting elements in a direct and intuitive way, they can be used to quantitatively understand their relationships. Nonetheless, it should be underlined that, even though graph modelling highlights the presence of interactions among elements, it partly loses the features characterizing the latter. For example, in modelling the complicated infrastructure of Internet, computers are represented simply as nodes linked by edges, though they are a complex intertwining of hardware and software. Nevertheless, the graph model can be enriched by assigning *attributes* to both nodes and edges. back to the Internet example, each edge may be assigned a numerical attribute describing the speed of data exchange along that connection while each node (i.e. computer) may be described by another numerical feature representing its speed in elaborating data.

Moreover, each node can be associated with its own measure of *importance* within the network. This importance is measured by the so-called *network metrics*. Different measures represent different types of node importance and, above all, provide information that cannot be derived from conventional statistical evaluations of databases but rather enrich the understanding of the Complex System under consideration.

Graph theory will be deepened in the next chapter.

1.3 Exploiting Complex System Machine Learning models

As described earlier, networks are the primary mathematical tools used to model all those systems composed of elements with mutual interactions like social networks [39]. The first step in studying Complex Systems is to identify the best network model. Determining the elements to be identified as nodes and the relationships to represent. Naturally, this step heavily depends on the considered dataset. Moreover, as explained previously, it is possible to enrich the network

model with additional data, if present in the dataset, providing attributes to both nodes and links. For instance, if the nodes of the network represent users of a Social Network, the attributes to associate with them can include age, geographical origin, educational level.

Furthermore, nodes can be even characterized by their importance in the network. There are different quantitative definitions of network importance and are obtained by using different network measures that quantify the value of nodes in the Complex System of which they are part. The quantitative definition of these measures will be presented in the following chapter. The added value of these metrics lies in being not obtainable by classical statistical analysis, since they take care of the network structure, i.e. the presence of relations among items. It is evident that the network metrics should be carefully chosen, depending on the task to be carried on using graph theory.

After deriving significant network centralities, it can be determined how they are related to the quantitative information contained in the dataset, e.g. buying habits of a consumer. In this way, it is possible to verify whether and how the importance of a node in the social network influences its characteristics as a user. This result can be obtained by using both *unsupervised* and *supervised machine learning models* [41, 42]. Unsupervised Machine Learning aims at discovering hidden patterns or grouping in the data without the need for human intervention [43]. One important example is *clustering* of data with algorithms like K-Means, K-Medoids or Hierarchical Clustering [44, 45, 46]. Supervised Machine Learning applied to network data results in *community detection*: finding sets of similar nodes relying on the network structure. A community is defined as a set of nodes having more links within the community itself than with nodes outside it. This definition is not mathematically well-posed, so that different community detection algorithms have been proposed [48, 49, 50].

Besides discovering hidden patterns or groups of tightly linked nodes, network information represented by network metrics can be used to build a model predicting the future characteristics of nodes or their future behavior. These tasks are accomplished using supervised Machine Learning tools [51]. These algorithms are based on a clear subdivision of data into *input* and *output* data. Input data are the features characterizing the samples fed into algorithms, and output data are the samples' features to be forecast. These algorithms start from a set of *training* samples, that are used to find the characteristics of the function linking input and output. This function can be linear or non-linear. Then, this function is used to forecast the output on *unseen data*, i.e. data not used for training. These data are called *test samples*. It should be noted that the procedure of dividing data into a training-set and a test-set must be accurate [53]. Accordingly, performance measures are used to quantify how good the algorithm is in generalizing training data to test on them. There are different performance measures for this purpose [54]. The main Machine Learning methods and algorithms will be thoroughly shown in the next chapter.

1.4 Thesis organization

In this chapter the links between Industry 4.0 and Complex Systems have been highlighted, showing how the latter are pivotal to fully describe and interpret the Fourth Industrial Revolution. Then, the two main pillars needed to extract

useful insights from Complex Systems have been introduced. Graph theory and Machine Learning. These topics are dealt with in depth in the following chapter.

In the third chapter three datasets are introduced on which graph modelling and Machine Learning algorithms are applied. One of them deal with an important Complex System in the context of Industry 4.0, at different levels of detail: the *startup ecosystem* (or *innovation ecosystem*). This ecosystem is defined as the set of startups, their investors and the corresponding funding relations. This is an example of Complex System that is often overlooked in its relation with Industry 4.0. The combined use of graph theory and Machine Learning algorithms will shed light on strategic elements in this system and on the correct evaluation of countries' innovation ecosystems.

The third dataset contains tourists' reviews about accommodation facilities in Apulia, a region in the South-East of Italy, with a strong tourism vocation. In this case Machine Learning will be used to analyze review textual data and highlight strengths and weaknesses of the Apulian tourism. This analysis could benefit both tourists and facilities' owners.

The fourth chapter shows how Complex Networks and Machine Learning have been effectively deployed in these three cases. The fifth chapter draws conclusions and future perspectives of this work.

Chapter 2

How to extract insights for Industry 4.0 graph theory and Machine Learning

Complex Systems, as highlighted in the previous chapter, are an invaluable tool to quantitatively understand and interpret the peculiar phenomena underpinning the Fourth Industrial Revolution like the ever-closer connection between firms and consumers. In this chapter, the main mathematical tools used to model and forecast the evolution of Complex Systems will be shown. In particular: the first section will deal with graph theory, the second section will discuss the main Machine Learning algorithm which can be used to exploit graph modelling to extract information about Complex Systems and forecast their evolution.

2.1 Fundamentals of graph theory

This chapter will first present some notable examples of Complex Systems, then the mathematical tool used to model these systems will be introduced: graph theory. After underlying its importance, some of the fundamental quantities that characterise networks will be discussed.

2.1.1 Complex Systems and graphs

As underlined in the previous chapter, a Complex System is a set of elements endowed with mutual relations. A great number of systems fall within this definition:

- *Internet* : a set of computers that are spread all over the world and linked by wires or electromagnetic connections that allow their communication and data transfer.
- *World Wide Web*: documents endowed with hyperlinks that allow the navigation from one document to another.

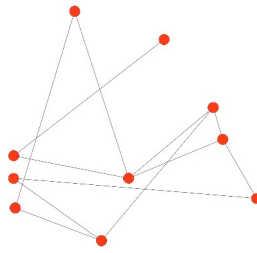


Figure 2.1. Graphical representation of an undirected graph.

- *Social network* : group of people together with their social relationships (e.g. friendship, kinship)
- *Neural network* : a set of brain cells (neurons) and their synapses that allow the passing of electromagnetic signals from one neuron to another one.
- *Metabolic network* : A complex of metabolites (substances such as carbohydrates, lipids, amino acids and nucleotides), which are linked by chemical reactions enabling the transformation of one metabolite into another.
- *Power grid* : Set of current generators, stations and substations for distributing electricity to consumers, which are connected by high-voltage lines, or transformers.
- *Stock market* : Set of shares subject to buying and selling on the stock exchange, for which the links are represented by coupled fluctuations.

It should be noted that among the previous examples there are some in which the links allow connection from one sub-system to another without any preference of direction (the Internet, the stock market, social and neural networks), while others are endowed with links that only allow connection from one sub-system to another but not vice versa (the World Wide Web and the power grids) the former have *undirected links*, while the latter present *directed links*.

Notwithstanding the difference among the previous systems they all can be modelled by a single mathematical *graph theory*. A graph, G , is a couple (V, L) where V is a non-empty set and L is a set of couples of elements of V [40]. The elements of V are called *vertices* (or *nodes*) while the couples in L are called *links* (or *edges*). Graph's definition allows a natural graphical representation where nodes are represented as points and their relations are depicted as lines joining the corresponding points. Moreover, while *undirected links* can be represented simply as lines, *directed links* are depicted as arrows beginning from the *source node* and ending in the *target node*. Figures 2.1 and 2.2 present an unirected and directed graph, respectively.

In the next section it will be shown the main reasons behind graphs' success in modelling Complex Systems.

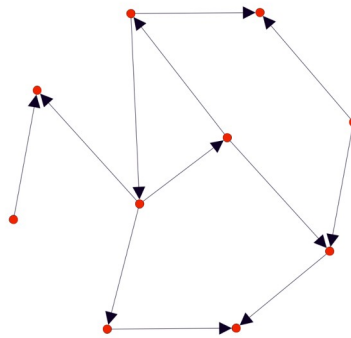


Figure 2.2. Graphical representation of a directed graph.

2.1.2 Graphs' value in modelling Complex Systems

The representation of Complex System by means of graph entails a simplification of the elements and their connections. For example, the computers of the Internet are represented by points, completely neglecting their internal structure (software and hardware), just as connections are depicted as line segments, ignoring their nature (wired, Wi-Fi, routers, etc.). This may lead one to consider the graphical representation of Complex Systems as unsuitable for studying them since some fundamental characteristics seem to be lost. Actually, the study by means of graph is an alternative to both the analysis of the individual network components and that of the nature of the connections, but allows another aspect to be studied: the pattern of connections. In other words, studying a Complex Network by means of a graph makes it possible to study how the elements are connected to each other. This type of study is not trivial because the way in which the connections are arranged influences the functioning of the system itself in terms of:

- its robustness or fragility with respect to the disappearance of its vertices;
- the rate of information passing through the system.

In this regard, it is interesting to consider the example of a system of computers shown as a graph in figure 2.3.

Communication is conveyed via a single terminal (*Computer 1*): in such a case, the connection between any two computers is guaranteed and passes only through *Computer 1*. If this element fails (e.g. because it is under a hacker attack), then computers would be unable to communicate with each other. In other words, the system is robust for random hacker attacks, because if a virus hits an arbitrary computer, communication between all the others is not affected (unless it is *Computer 1*), but it is extremely susceptible to attacks targeted at *Computer 1*. Moreover, it must also be added that the amount of work involved in coordinating communication would only burden *Computer 1*.

From the point of view of the representation of systems, it is however possible to increase the information contained in graphs by enriching vertices and/or links with *attributes*. For example:

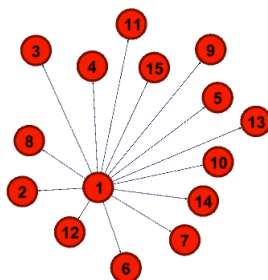


Figure 2.3: Example of a set of computers whose communication is conveyed by a single computer (Computer 1).

- In the case of Internet, it is possible to associate the data transfer rate to the links, or the average amount of data transferred in the unit of time, by individual connections;
- As regards social networks, it is possible to endow links with an attribute indicating the type of relationship that binds a pair of individuals: friendship, kinship, hatred, indifference.
- Considering the World Wide Web, it is possible to associate vertices with the number of visitors in a certain period while oriented links can be assigned information on the number of visitors clicking on that link.

It is obvious that the number and type of attributes to associate with links and vertices depend on the particular information to be emphasised.

From the definition and examples given in the previous section, the pervasiveness of the notion of Complex System is evident. In this reason it is not surprising that, until a few decades ago, there was no single corpus of notions and methods for studying them. In this regard, two early outstanding milestones dealing with the mathematical methods of graph theory are [56] and [57]. Actually, their knowledge was dispersed among the various branches like social sciences and biology. On the one hand, this represented an obstacle in the advancement of the study of Complex Systems (for example, methods of network analysis, known to social scientists, were not known to biologists and chemists) but, on the other hand, it generated a considerable wealth of viewpoints, notions and methods of analysis [39].

2.1.3 Multi-graph, simple graph and weighted graph

In addition to the already mentioned *directed* and *undirected graphs*, there are various types of graphs that are useful in different situations. Among the most important there are the *multigraph* and the *weighted graph*.

A *multigraph* is a (directed or undirected) graph in which at least one couple of nodes is linked by more than one link. These couples are indicated as linked by a *multilink*. An example of multigraph can be observed in figure 2.4

A *weighted graph* is a (directed or undirected) graph whose links are endowed with a numerical attribute (e.g. every link is associated with a real number). This attribute is called *weight* of the link.

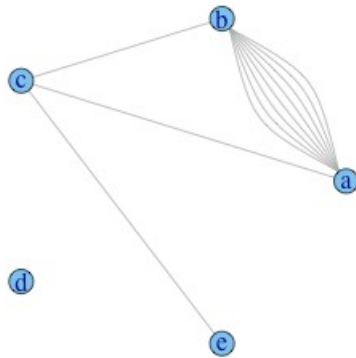


Figure 2.4 Example of a multigraph couple (a,b) is linked by a multilink.

For example Internet with average data rate values associated with each connection can be modelled as a weighted graph.

If a link (in a directed or undirected graph) begins and ends on the same vertex, then that link is called *selflink*. A graph that contains neither selflink nor multilink is called *simple graph*.

2.1.4 Degree and Adjacency matrix

Once the definition of a graph and its main characteristics have been established, it is useful to consider some other notions that are fundamental for complex Systems analysis. The most immediate and important is that of *degree* of a vertex. This definition has different expressions depending on whether a directed or undirected graph is considered.

For an undirected graph, the *degree* of a vertex is defined as the number of links attached to the vertex. For example, considering figure 2.2, vertex 1 has degree 14 while all the others have degree equal to 1.

In case of directed graphs there are two possible definitions of *degree*: *degree* and *outdegree*. The *indegree* of a vertex is defined as the number of edges entering the vertex (those having the vertex as target node). On the contrary, the *outdegree* of a vertex is defined as the number of its outgoing links.

The *Adjacency matrix* is a mathematical representation of graph, and its form depends on the considered type of graph. In particular, if the graph has n vertices and we arbitrarily assign unique numeric indices from 1 to n to the vertices, the Adjacency matrix is a $n \times n$ matrix with generic element A_{ij} ($i, j = 1, \dots, n$) defined in the following manner (assuming the absence of *selflink*):

- Simple undirected graph

$$A_{ij} = \begin{cases} 1 & \text{if node } i \text{ is linked to node } j \\ 0 & \text{otherwise} \end{cases}$$

- Undirected multigraph

$$A_{ij} = \begin{cases} m & \text{If } m \text{ links are present between } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

- Weighted undirected graph

$$A_{ij} = \begin{cases} r & \text{if a weighted link is present between } i \text{ and } j \text{ having weight } r \\ 0 & \text{otherwise} \end{cases}$$

- Simple directed graph

$$A_{ij} = \begin{cases} 1 & \text{if there is a link having } i \text{ as source node and } j \text{ as target node} \\ 0 & \text{otherwise} \end{cases}$$

- Directed multigraph

$$A_{ij} = \begin{cases} m & \text{if there are } m \text{ links having } i \text{ as source node and } j \text{ as target node} \\ 0 & \text{otherwise} \end{cases}$$

- Weighted directed graph

$$A_{ij} = \begin{cases} r & \text{if there is a link having } i \text{ as source node and } j \text{ as target node having weight } r \\ 0 & \text{otherwise} \end{cases}$$

It is evident that the Adjacency matrix is *symmetric* for undirected graphs, whereas in general it is not for directed graphs. In the absence of selflinks the diagonal terms of the previous matrices are all zeros. On the other hand, selflinks are present, their presence is indicated by diagonal elements:

- Undirected graph with selflinks

$$A_{ii} = \begin{cases} 2p & \text{if } p \text{ selflinks are present at node } i \\ 0 & \text{otherwise} \end{cases}$$

- Weighted undirected graph with selflinks

$$A_{ii} = \begin{cases} \sum_{k=1}^p r_k & \text{if } p \text{ selflinks are present at node } i \text{ with weights } r_k \\ 0 & \text{otherwise} \end{cases}$$

- Simple directed graph with selflinks

$$A_{ii} = \begin{cases} p & \text{if } p \text{ directed selflinks are present at } i \\ 0 & \text{otherwise} \end{cases}$$

- Weighted directed graph with selflinks

$$A_{ii} = \begin{cases} \sum_{k=1}^p r_k & \text{if } p \text{ selflinks are present at node } i \text{ with weights } r_k \\ 0 & \text{otherwise} \end{cases}$$

There is a very close relationship between vertices in a graph and the relative Adjacency matrix. Indeed, from the definitions above, it is evident that, in an undirect graph with n vertices, k_i denotes the degree of the vertex i , the following formula holds:

$$k_i = \sum_{j=1}^n A_{ij} . \quad (2.1)$$

Analogously, for a directed graph:

$$k_i^{out} = \sum_{j=1}^n A_{ji} . \quad (2.2)$$

$$k_i^{in} = \sum_{j=1}^n A_{ij} . \quad (2.3)$$

It is important to emphasise that the degree of a vertex, representing its number of links, is an immediate measure of its *importance* within the graph. In fact, it is natural to consider a vertex as more influential the higher the number of links it has. On the other hand, however, this observation can be countered by asserting that in many cases the number of links alone cannot represent a complete measure of the vertex's importance, but the importance of the vertices to which it is connected must also be considered. In this regard, consider two elements (vertices) of a Social network, denoted as A and B , such that $k_A > k_B$. This means that A knows more people than B . Therefore, considering only the degree, A is more important than B . Now, suppose that B knows only *Socially Influential* peoplesuch as academics and political authoritieswhile A knows none of these. To what extent is it still possible to argue that B is less important than A ? It is obvious that other measures of importance must be determined to account for this observation. These will be discussed in the next sections.

2.1.5 Geodetic paths

This section contains the fundamental notion of *geodetic path*, which will allow the introduction of new *network centralities* different from the *degree*. In this regard, the following definitions of *paths* are given. Given an undirected graph and an arbitrary couple of nodes i and j , a *path* between i and j is defined as a sequence of adjacent nodes (i.e. directly linked by edges) together with their linking edges that begins with node i (resp., node j) and ends with node j (resp., node i). If no path can be determined between i and j , then these nodes are defined as *disconnected*. As regards directed graphs, because of the orientation of the links, an *oriented path* can be defined. Given a directed graph and an arbitrary pair of vertices i and j , an *oriented path* between i and j is a sequence of adjacent nodes and the corresponding oriented links connecting them consecutively in pairs, beginning at i and ending at j . Figures 2.5 and 2.6 show two examples of paths in the case of undirected and directed graphs, respectively.

The first feature that can be used to describe a path is its *length*, equal to the number of links of the path. This definition is valid for both directed

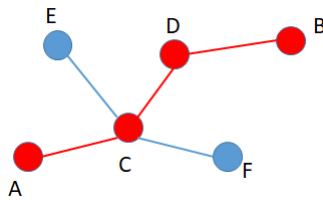


Figure 2.5 Example of a path (in red), between node A and B, in an undirected graph.

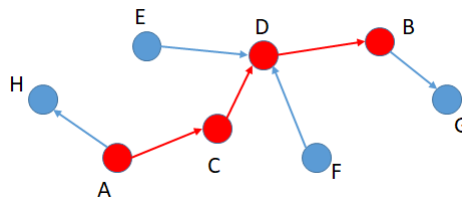


Figure 2.6 Example of a path (in red), between node A and B, in an undirected graph.

and undirected graphs. For example, both the paths in figures 2.5 and 2.6 have length equal to 3.

The notion of path in a graph forms the basis for the notion of *geodetic path*. Given a graph (directed or undirected) and a pair of vertices i and j , a *geodetic path* between i and j is defined as the path with the smallest length among all the possible paths between i and j . If no paths exist between i and j , then their geodetic length is set equal to ∞ .

As a final remark it is interesting to relate the number of paths of a given length within a graph (directed or undirected) to appropriate powers of the corresponding Adjacency matrix. To determine this relationship, consider for example, a simple graph with n vertices and a couple (i, j) of vertices. If i and j there is a path of length 2 if there is a third vertex k , such that i is connected to k and k to j . More formally, we can say that such a path is present if

$$A_{ki} A_{jk} = 1 \tag{2.4}$$

otherwise $A_{ki} A_{jk} = 0$.

If the number of paths between i and j is denoted by n_{ij} , then, from equation 2.4, it can be derived that:

$$n_{ij} = \sum_{k=1}^n A_{ki} A_{jk} = (\mathbf{A}^2)_{ji}$$

Analogously, the number of paths having length $r > 0$ between nodes i and j , $n_{ij}^{(r)}$, is equal to:

$$n_{ij}^{(r)} = (\mathbf{A}^r)_{ji} \tag{2.5}$$

where \mathbf{A}^r is the r -th power of the Adjacency matrix of the considered graph.

2.1.6 Closeness

In this section, we will consider a type of vertex centrality, the *closeness*, which is not based on the notion of degree. In fact, *closeness* is defined by means of the geodesic paths seen in the previous section. In particular, it is necessary to define the notion of *average length* as follows: Given a graph (directed or undirected) with n vertices and an arbitrary node, i , if d_{ij} is the length of the geodesic path between i and another node j , then the *average length* relative to node i is defined as:

$$l_i = \frac{1}{n} \sum_{j=1}^n d_{ij} \quad (2.6)$$

From this definition, it is evident that the closer l_i is to 1 (the minimum value for l_i , corresponding to the vertex i directly connected to all the other vertices of the graph), the closer i is, on average, to all the other vertices of the graph. In such a case, depending on the represented system, i can more easily influence (or be influenced by) the other vertices, be more readily reached by information from the other vertices of the graph, or disclose its ideas or opinions more easily. Accordingly, the smaller l_i the more important i can be considered, and a measure of such importance must be related to the *inverse* of l_i . In particular, for an undirected graph, can be defined the *closeness* of node i by the following equation:

$$C_i = \frac{1}{l_i} = \frac{1}{n} \sum_{j=1}^n \frac{1}{d_{ij}} \quad (2.7)$$

Notwithstanding its simplicity, this definition has two issues:

- From a mathematical point of view, if vertex i and vertex j are not connected by convention $d_{ij} = \infty$. This means that, unless the vertices of the graph are all connected by at least one path (in this case the graph is called *connected*) the closeness is null. Since most networks encountered in applications are not connected, it is evident that this type of centrality would be of little use;
- When applied to graphs modelling real complex systems, closeness has a limited range of values. The minimum and maximum values differ by a term of order 10^{-5} , so that it is difficult, in general, to identify the vertices with greater centrality. In other words, *closeness* is very sensitive to the graph structure, so any updating of it can lead to a drastic change in *closeness* [39].

The last problem cannot be solved, since it is intrinsic in the definition of the closeness centrality. In fact, most graphs in applications have a *diameter* (i.e. the largest geodesic distance among the finite ones) of order \sqrt{n} , where n is the number of vertices of the graph, which is a value varying in the range 1 – 10: the diameter represents the upper limit of the values of d_{ij} , while the lower limit is 1. It is evident, therefore, that the values of d_{ij} (and hence of l_i) vary over a very limited range.

The first problem is solved by slightly modifying the definition of closeness in the following manner:

$$C_i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{1}{d_{ij}} \quad (2.8)$$

The (2.8) is the *harmonic mean* of the values of geodesic distance between vertices, and it is evident that the terms d_{ii} (relative to unconnected pairs of vertices) *do not* contribute to the value of C_i . Solving the first problem in addition, in the summation of equation (2.8) the term $j = i$ is deleted, because it would give $d_{ii} = 0$ and thus $\frac{1}{d_{ii}} = \infty$. This accounts for the term $\frac{1}{n-1}$ instead of $\frac{1}{n}$.

For the sake of completeness, it should be pointed out that the (2.8) is rarely used in applications; equation (2.7) is typically used, which, instead of the term $d_{ii} = \infty$ one substitutes the term n , equal to the number of vertices, which is typically a much higher value than the other finite terms, thus *simulating* the infinite term.

What has been said so far refers to the case of an undirected graph, but it is possible to express the same notions for directed ones, taking into account the orientation of the path. Given a directed graph and an arbitrary vertex i , the *incloseness* of vertex i is defined as:

$$C_i^{(in)} = \frac{1}{n} \sum_{j=1}^n \frac{1}{d_{ij}^{(in)}} \quad (2.9)$$

where $d_{ij}^{(in)}$ is the length of the geodesic path going from j to i . Analogously, the *outcloseness* of an arbitrary vertex i of a directed graph is defined as follows:

$$C_i^{(out)} = \frac{1}{n} \sum_{j=1}^n \frac{1}{d_{ij}^{(out)}} \quad (2.10)$$

where $d_{ij}^{(out)}$ is the length of the geodesic path going from i to j .

It seems wise to underline that the relationship between closeness and degree centrality is thoroughly studied in [58].

2.1.7 Betweenness

In this section, a further notion of vertex centrality will be defined which is based on the notion of paths in a graph. This centrality is called *betweenness* and is defined as follows.

Given a graph (directed or undirected) and an arbitrary vertex i , the *betweenness* of i is related to the number of geodesic paths between possible pairs of vertices of the graph passing through i . In more formal terms, if g_{st} represents the total number of geodesic paths existing between vertices s and t and n_{st}^i represents the number of such geodesic paths passing through vertex i , then the *betweenness* of i is:

$$B_i = \sum_{s,t} \frac{n_{st}^i}{g_{st}} \quad (2.11)$$

For undirected graphs the summation in equation (2.11) leads to a double counting of paths for the same pair of vertices, but this is not seen as a problem

because it is the relative importance between nodes that matters and not the absolute value of the centralities.

The interpretation of betweenness centrality for a vertex is simple: a vertex with high betweenness has a great deal of control over communications between elements in the graph (assuming that such communications occur through geodesic paths). Therefore it can be considered as an important vehicle of knowledge/influence between vertices. As an example, just think of Social networks of any kind: a vertex with high betweenness facilitates knowledge between a pair of other elements, which, through that vertex, can communicate and know each other.

2.2 Machine Learning algorithms

In this section the main Machine Learning algorithms used in this thesis will be exposed. According to section 1.3, these tools can be mainly subdivided in two categories: *unsupervised* and *supervised* algorithms. Unsupervised Machine Learning aims at discovering hidden patterns or groups in data, supervised algorithms are fed with some input features of data in order to forecast the corresponding output feature (categorical or numerical). First, supervised community detection algorithms for graphs are discussed, then the main supervised tools used in this work are shown.

2.2.1 Unsupervised Machine Learning community detection in graphs

A common task in graphs applications is that of *community detection*: the search for the naturally occurring groups in a graph regardless of their number or size. This is a tool for discovering and understanding the large-scale structure of graphs and, as a consequence, of the Complex Systems these networks are modelling. In particular, community detection aims at finding a natural subdivision of nodes in groups such that there are more links within groups than among them. These groups are called *communities* of nodes. This is a mathematically ill-posed problem since there is no a unique definition of *natural* subdivision of nodes. Accordingly, many algorithms have been defined to accomplish community detection based on different definitions of a *natural subdivision* [32]. An example of communities in a graph is shown in figure 2.7

The most successful algorithms are based on the maximization of a particular function called *modularity*. These algorithms are based on the following consideration: if we find a partition of nodes that has few edges between its groups, but the number of such edges is about what we would have expected were edges simply placed at random in the graph, then this nodes' subdivision would hardly be defined significant.

As a matter of fact, in the conventional development of this idea one considers not the number of edges between groups but the number within groups. It should be noted that the two approaches are equivalent: the number of edges that lies within a community necessarily does not lie between groups. The number can be calculated from the other given the total number of edges in the graph. Therefore the goal will be to find a measure that quantifies how many edges lie within groups in our network relative to the number of edges

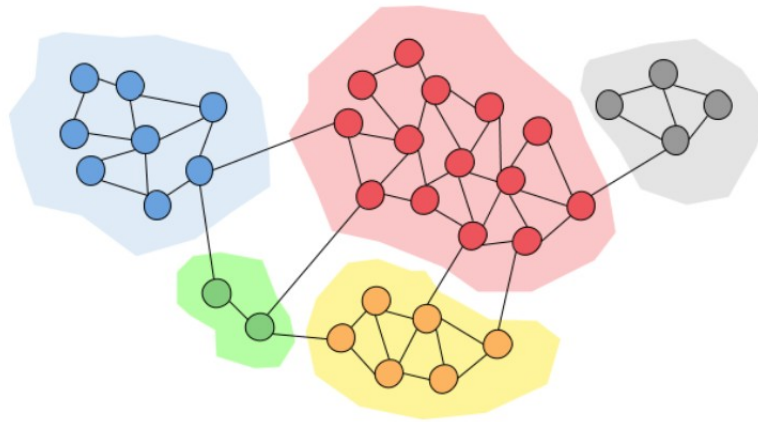


Figure 2.7: Communities in a graph Source:Thamindu Dilshan Jayawickrama, *Example of communities in a graph. These communities have been found through modularity maximization,* image file 2021, Towards Data Science, PNG, <https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae> (last accessed October 2023).

expected on the basis of chance. This measure is the modularity [32, 50]. From a mathematical point of view, it is defined as follows (for an undirected simple graph):

$$Q = \frac{1}{2m} \sum_{ij} A_{ij} - \frac{k_i k_j}{2m} \delta(c_i c_j) \quad (2.12)$$

where: A_{ij} is the (i, j) term of the Adjacency matrix; k_i (resp. k_j) is the degree of node i (resp. node j); m is the number of edges in the graph; c_i (resp. c_j) is the community to which node i (resp. node j) belongs and $\delta(c_i c_j)$ is the Kronecker delta between these groups: $\delta(c_i c_j) = 0$ if $c_i \neq c_j$ while $\delta(c_i c_j) = 1$ if $c_i = c_j$. It should be noted that the term $\frac{k_i k_j}{2m}$ is the probability of having a random link between node i and node j while A_{ij} is 1 if these nodes are linked and 0 otherwise. The term $A_{ij} - \frac{k_i k_j}{2m}$ thus represents the difference between the number of edges linking two arbitrary nodes of the graph, i and j , and those that would be expected on the basis of chance.

It should be noted that modularity can be seen as a measure of *graph assortativity*. *Assortativity* indicates the tendency of similar nodes to be linked with each other, where the *similarity* of nodes can be defined on the basis of categorical or scalar attributes. In general, it is well known that social links (e.g. acquaintances, business relations) are created on the basis of similar attributes like age, income or even race [39]. In this case, graphs are denoted as *assortative*. On the contrary, when edges are formed between dissimilar nodes, the graph is referred to as *disassortative*. One particular case of assortativity arises when the node attribute to be considered is the *degree*. This represents a property of the network structure. In particular, *degree-assortativity*, is measured by the following equation [39]:

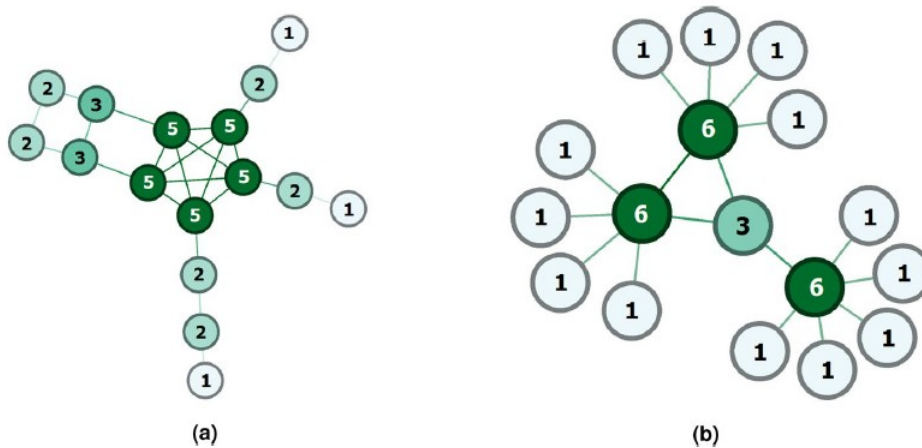


Figure 2.8: Examples of assortative and disassortative graphs. High-degree nodes are colored dark green while low-degree nodes are colored with a lighter color. Each node is labelled with its degree. (a) Assortative graph ($r = 0.60$) where high-degree nodes are attached to high-degree nodes and low-degree nodes are attached to low-degree nodes. (b) Disassortative graph ($r = 0.84$) where high-degree nodes are attached to low-degree nodes. Source: [60].

$$r = \frac{1}{2m} \sum_{ij} A_{ij} - \frac{k_i k_j}{2m} \quad (2.13)$$

As in formula 4.3 the term in parentheses in equation 2.13 measures the difference between the number of edges with similar degree values and that expected on the basis of chance. A degree-assortative graph is characterized by $r > 0$ and represents a situation in which high-degree nodes (core nodes) are connected with themselves, so producing a *core/periphery* structure, as shown in figure 2.8 (a). The core is composed by high degree nodes connected with themselves, while the low-degree vertices represent the periphery. Contrary, degree-disassortative graphs are characterized by $r < 0$ and have a *star-like* structure, in which high-degree nodes are linked to a high number of low-degree vertices, as shown in figure 2.8 (b). For example, the graph in figure 2.3 shows a clear star-like structure, in which *Computer 1* has the highest degree while all the other terminals are linked only to it. Accordingly, it can be considered as a degree-disassortative graph.

Going back to modularity, it measures the assortativity of nodes belonging to the same community, which represents a categorical attribute.

Accordingly, one way to detect communities in networks is to look for the divisions that have the highest modularity and, in fact, this is the most commonly used method for community detection [61].

There is a great variety of algorithms for maximizing (or minimizing) functions over sets of states and anyone of them could be used for the modularity maximization problem, thereby creating a new community detection algorithm [62, 63, 64, 65, 50, 66]. Each of them seeks to maximize modularity over divisions into any number of communities of any sizes and thus to determine

both the number and size of communities.

The first and one of the most widely used optimization strategies is *simulated annealing*, which exploits the physics of slow cooling or "annealing" of solids [62, 64, 63]. In brief, it is known that a hot system, such as a melted metal, will, if cooled sufficiently slowly to a low enough temperature, eventually find its ground state: the state of the system that has the lowest possible energy. The algorithm of simulated annealing works by treating the quantity of interest, modularity, as an energy and then simulating the cooling process until the system finds the state with the lowest energy. Since the goal is finding the highest modularity, energy is equated to minus the modularity. The main disadvantage of the approach is that it is slow, typically taking several times as long to reach an answer as competing methods do.

Another method makes use of the so-called *greedy algorithm*. In this approach every vertex is initially associated to a one-vertex group of its own and then pairs of groups are amalgamated in successive steps. The groups to be merged, at each step, are those whose joining gives the biggest increase in modularity or the smallest decrease if no choice gives an increase. The process continues until all vertices are amalgamated into a single large community. Among all the states through which the graph passed during the course of the algorithm, the one with the highest modularity is chosen. The modularity values achieved by this method are in general somewhat lower than those found by the simulated annealing method. But, on the other hand, this is one of the few algorithms fast enough to work on the very largest networks now being explored [67, 68].

Besides modularity maximization algorithms, another class of methods is that of *Hierarchical Clustering*. The goal of these algorithms is to find communities in graphs through a hierarchical decomposition into a set of nested communities, rather than just a single division into a single set of communities. Usually, these nested communities are shown in the form of dendrograms, as depicted in figure 2.9.

Precisely Hierarchical Clustering defines a set of agglomerative techniques in which the initial status has the individual vertices as groups on their own and are iteratively joined together to form larger groups. Even though the previous greedy modularity maximization algorithm is an example of agglomerative method, Hierarchical Clustering methods generalize this approach. In particular, Hierarchical Clustering is based on the definition of a measure of similarity between vertices, based on the graph structure, and then on the merging of the closest or most similar vertices to form groups. There can be defined many suitable measures of nodes' similarity [32].

This freedom in the choice of similarity measures is both a strength and a weakness of the Hierarchical Clustering method. In fact, it allows this method to be tailored to specific problems, but it also means that it may give different answers depending on the chosen similarity measure. This also means that there is no way to know if one measure will yield more useful information than another. As a consequence, the choice of the similarity measure is determined more by experiment than by first principles.

Once a similarity measure is chosen, it must be calculated for all pairs of vertices in the graph. Then, those vertices having the highest similarities must be merged. This, however, leads to the following problem: the similarities can give conflicting messages about which vertices should be grouped. For example,

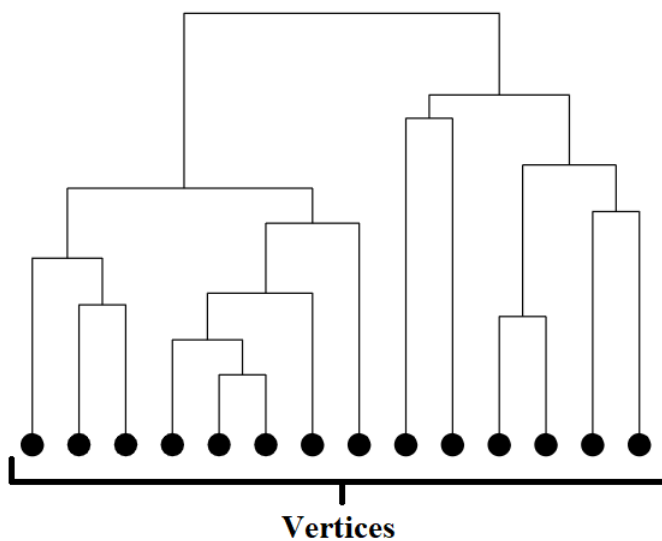


Figure 2.9: Example of a dendrogram produced by a Hierarchical Clustering algorithm.

if vertices A and B have high similarity, as do vertices B and C, it may be argued that A, B, and C should all be in a group together. But what happens if A and C have a low similarity? Should A and C be in the same group or not?

In order to solve this issue, Hierarchical Clustering methods define a measure of similarity between *groups of nodes* based on node similarity. This groups' similarity measure is built combining that among vertices.

There are three common ways of combining vertex similarities to give similarity scores for groups. They are called *single-linkage*, *complete-linkage*, and *average-linkage* clustering methods. In the single-linkage clustering method, similarity between the two groups is defined to be the highest similarity between all possible couples of nodes, where one vertex comes from one group and the other from the second group. Accordingly, only a single vertex pair needs have high similarity for the groups themselves to be considered similar.

At the other extreme, complete-linkage clustering defines the similarity between two groups to be the similarity of the least similar pair of vertices. In contrast with single-linkage clustering this is a very stringent definition of group similarity: every single vertex pair must have high similarity for the groups to have high similarity.

In between these two extremes lies average-linkage clustering. In this method, the similarity of two groups is defined to be the mean similarity of pairs of vertices.

Then, the full hierarchical clustering method is as follows:

1. Choosing a similarity measure and evaluate it for all vertex pairs.
2. Assigning each vertex to a group of its own, consisting of just that vertex. Then, the initial similarities of the groups are simply the similarities of the vertices.

3. Find the pair of groups with the highest similarity and merge them into a single group.
4. Calculate the similarity between the new composite group and all others using single-, complete-, or average-linkage.
5. Repeat from step 3 until all vertices have been joined into a single cluster.

2.2.2 Supervised Machine Learning algorithms and Explainability

As explained in section 1.5, supervised machine learning is defined by its use of labelled datasets to train algorithms to classify data or predict outcomes accurately (denoted as *regression problems*). Supervised learning helps organizations solve a variety of real-world problems, such as classifying spam in a separate folder from your inbox or fraud-detection [69, 70]. In the next section a statistically robust framework to train models will be presented. Then, the main classification algorithms will be presented. After, the classification metrics used to quantify models' performance will be described. The final section will deal with a hot topic in Machine Learning: Explainability tools used to make model decisions more transparent to experimenters and users.

Strength the training: the cross-validation framework

As train input data are fed into the model, the latter adjusts its parameters until the model has been fitted appropriately. This process is called *training phase* of the model. The training can be done in different ways, in order to ensure the statistical robustness.

In fact, learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called *overfitting*. To avoid it, it is common practice when performing a supervised machine learning experiment to hold out part of the available data as a test set.

When evaluating different settings of the models' parameters there is still a risk of overfitting on the test set because the parameters can be tweaked until the estimator performs optimally. As a result, knowledge about the test set can leak into the model so biasing the evaluation of the generalization performance. To solve this problem, yet another part of the dataset can be held out as a so-called *validation set*: training proceeds on the training set, after which evaluation is done on the validation set, and when the experiment seems to be successful, final evaluation can be done on the test set.

Nonetheless, by partitioning the available data into three sets, the number of samples which can be used for learning the model is drastically reduced. Moreover, the results can depend on a particular random choice for the pair of (train, validation) sets.

A solution to this problem is a procedure called *cross-validation (CV)*. The basic approach, called *k-fold CV*, the training set is split into k smaller sets, called *folds*. The following procedure is followed:

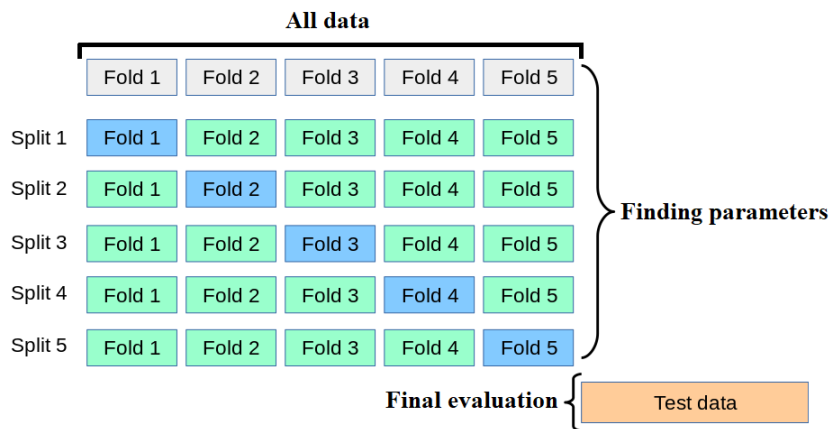


Figure 2.10: Graphical representation of k-fold cross-validation framework with k=5. Source: Scikit-learn development and maintenance team, file, https://scikit-learn.org/stable/_images/grid_search_cross_validation.png (last accessed 2nd October 2023).

1. a model is trained using $k - 1$ folds as training data;
2. the resulting model is validated on the remaining part of the data.

This procedure is graphically reported in figure 2.10

The performance measure reported by k-fold CV is then the average of the values computed in the loop. This approach can be computationally expensive, but does not waste too much data, as is the case when fixing an arbitrary validation set.

Main classification algorithms

The oldest and most commonly used model used in classification problems is the *Logistic Regression* [71,72]. It aims at modelling the probability of an event taking place by using the linear combination of one or more independent variables. If the classes to be predicted are two, then it is called *binary logistic regression* and the single binary dependent variable (the output) is coded by an indicator variable taking values 0 and 1. Formally, if X_1, \dots, X_N are the input features characterizing a sample (data), the probability of it being labelled with $y = 1$, according to the Logistic Regression model, has the form:

$$p(y = 1) = \frac{e^{a_1 X_1 + \dots + a_n X_n}}{1 + e^{a_1 X_1 + \dots + a_n X_n}}$$

Moreover, $p(y = 0) = 1 - p(y = 1)$ holds.

Even though Logistic Regression is naturally defined for tackling problems with binary dependent variables, it may be used even in for *multi-class classification problems*, where the output of each data sample can take more than two values. In this case, there are two heuristic approaches that can be used: *One-versus-One (OvO)* and *One-versus-Residual (OvR)* [73]. Delving into these methods:

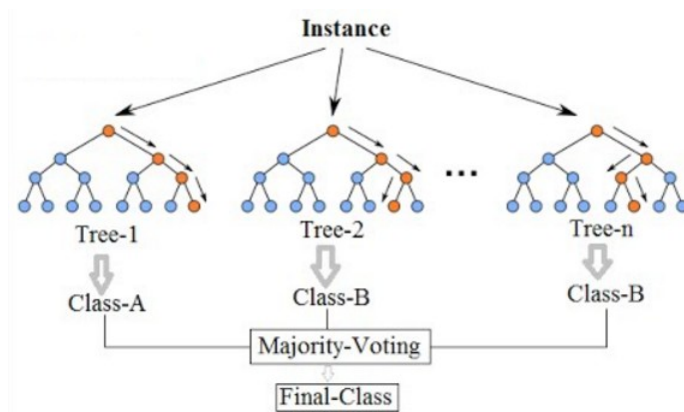


Figure 2.11 How a Random Forest determines its output from the trees in its ensemble [77].

- **One-versus-Rest.** It involves splitting the multi-class dataset into multiple binary classification problems. A binary classifier is then trained on each binary classification problem and predictions are made using the model that is the most confident.
- **One-versus-One.** Like OvR, OvO splits a multi-class classification dataset into binary classification problems. Unlike one-vs-rest that splits it into one binary dataset for each class, the one-vs-one approach splits the dataset into one dataset for each class versus every other class.

Another off-the-shelf classification algorithm is *Random Forest* (RF) [74]. It is a generalization of classical *decision trees* [75]. In fact, Random Forest is an ensemble learning method that works by constructing a multitude of trees during training. In particular, every tree is trained on a *bootstrapped* sample of training data (i.e. obtained by sampling with replacement from training data) and each tree uses a random subset of predictors to take decisions, in order to overcome the presence of strong predictors. The output of the Random Forest is the class selected by most trees, i.e. the so-called *majority vote rule*. Random Forest is built by merging different *base classifiers* (the decision trees) since decisions based on an ensemble of classifiers greatly improves the performance of a single decision tree [76]. It should be noted that Random Forest natively supports multi-class classification. Figure 2.11 best summarizes how a Random Forest works in classification settings.

A variant of Random Forest is the *Extreme Gradient Boosting* (XGB) classifier. In fact, XGB, like Random Forest, is a model in the form of an ensemble of decision trees, but, differently from Random Forest, it is built in an iterative fashion and its learning is slower than Random Forest. In particular, while trees in RF are trained on different bootstrapped samples taken from the training dataset, independently of each other, XGB does not involve any bootstrapping procedure but every tree is grown using information from previously grown trees, being fit on a modified version of the training data. The main idea underpinning XGB is that, given the current model, a decision tree is fit to the residuals from the current model. Then, this new decision tree is added

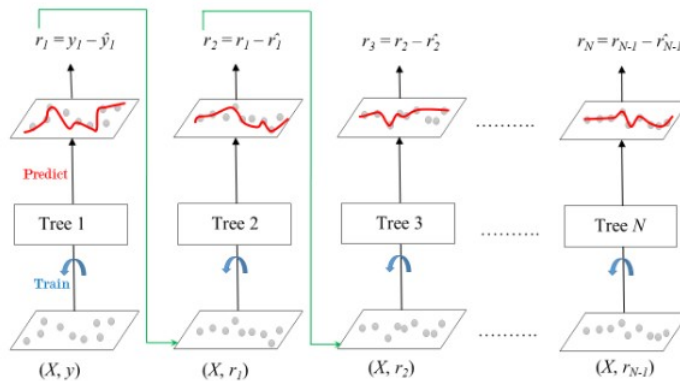


Figure 2.12 Training phases of a XGB classifier. Source: Geeks For Geeks development team, PNG file, <https://www.geeksforgeeks.org/ml-gradient-boosting/> (last accessed 2nd October 2023).

into the current model in order to update the residuals. By iteratively fitting trees to the residuals, the current model is improved on data where it does not perform well. Figure 2.12 shows the phases of XGB training.

It can be readily seen that while Logistic Regression provides a functional form of the probability, Random Forest and XGB does not. In this regard, the latter are more adaptive than Logistic Regression. Nonetheless this flexibility does not necessarily implies a better performance. This is particularly true in cases where a great number of classes are present [79]. Accordingly more biased models deserve attention. Together with Logistic Regression, another widely used biased model is the *Gaussian Naive Bayes* (GNB) classifier.

GNB classifier is based on Bayes' theorem and assumes some strong hypotheses about the independence of input variables in determining the probability of an item to belong to an output class [80].

In particular, if an instance determined by N input variables, (X_N) , should be assigned to one of K classes, (C, \mathcal{C}) , GNB aims at calculating the corresponding conditional probabilities $p(C_i | X_1, \dots, X_N)$, $\forall i \in \{1, \dots, K\}$. In order to determine these probabilities, GNB refers to Bayes' Theorem:

$$p(C_i | X_1, \dots, X_N) = \frac{p(X_1, \dots, X_N | C_i) p(C_i)}{p(X_1, \dots, X_N)} \quad (2.14)$$

where $p(C_i)$ is called *prior probability*; $p(X_1, \dots, X_N | C_i)$ is denoted as *likelihood distribution*; $p(X_1, \dots, X_N)$ is referred to as *evidence distribution*. GNB assumes that likelihood distributions are Gaussian, whose parameters should be estimated in the training phase of the model.

Since $p(X_1, \dots, X_N | C_i) p(C_i) = p(X_1, \dots, X_N, C_i)$, then applying simple probability rules and considering the hypothesis of mutual independence of the N input variables, the following formula holds

$$p(C_i | X_1, \dots, X_N) = \frac{p(C_i)}{p(X_1, \dots, X_N)} \prod_{j=1}^N p(X_j | C_i) \quad (2.15)$$

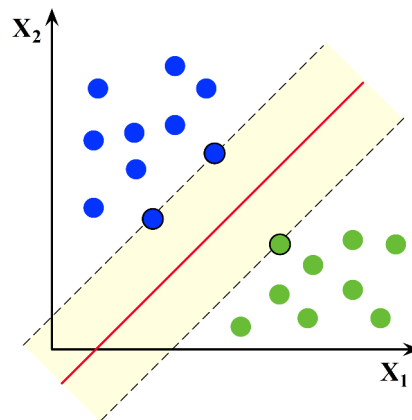


Figure 2.13: The result of the SVM algorithm applied to a dataset with two input features (X_1, X_2), for the sake of clarity. The two classes are reported in blue and green directly as colors of the data points. This maximum margin hyperplane is reported in red. Source: [83].

Finally, GNB will assign a class to every item \hat{y} if it has the greatest conditional probability. Mathematically:

$$\hat{y} = \max_{i \in \{1, \dots, K\}} p(C_i) \prod_{j=1}^N p(X_j | C_i) \quad (2.16)$$

Another kind of classification algorithm takes into consideration an embedding of data points in an Euclidean space, the so called *feature space*, in order to find an hyperplane able to distinguish points belonging to the different classes. In particular, every data sample, characterized by N input features, is represented as a point in the N -dimensional Euclidean space. This space is the *feature space*. The main algorithm in this field is the *Support Vector Machine* (SVM) [81]. SVM natively works for binary classification problems but, as the Logistic Regression, it can be deployed even for multi-class classification problems using the OvO and OvR frameworks. SVM is based on finding in the feature-space, the best hyperplane subdividing training data points of one class from those belonging to the other one. In particular, considering a training dataset of M items and with N input features, these items may be represented as $(X_1, y_1), \dots, (X_M, y_M)$, where X_i is the N -dimensional vector of input variables of i -th data item and y the corresponding binary label (0 or 1). They may be considered as geometric points in the N -dimensional feature space. The target of the SVM algorithm is to find the *maximum margin hyperplane*: the hyperplane which is defined so that the distance between the hyperplane and the nearest points from either group is maximized. These points are called *support vectors*. Figure 2.13 clearly explains the result of the SVM algorithm in a dataset with two input features.

Performance metrics for classification algorithms

According to the well-known *No Free Lunch Theorem*, it does not exist a model whose performance overcomes that of the others in all classification prob-

lems [84]. Accordingly, measuring algorithm performance is of paramount importance in every classification task. Moreover, there exist several different metrics highlighting different model behaviour [85]. The most widely used metrics, that are used even in this work, are the following:

- **Accuracy (acc).** It is defined as the ratio between correctly classified samples and the total number of samples. In formula:

$$acc = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP (True Positive) and TN (True Negative) are the correctly classified samples, while FP (False Positive) and FN (False Negative) are the wrongly classified samples.

- **Sensitivity (sens).** It is also called *Recall* or *True Positive Rate* and is defined as the ratio of the positive correctly classified samples.

$$sens = \frac{TP}{TP + FN}$$

- **Specificity (spec)** It is also denoted as *True Negative Rate* and is the ratio of the negative correctly classified samples.

$$spec = \frac{TN}{TN + FP}$$

- **F1-score (F1).** It is defined as the harmonic mean of correctly classified samples. In formula:

$$F1 = \frac{TP}{TP + \frac{FP + FN}{2}}$$

- **Area Under Curve - Receiver Operating Characteristic (AUC-ROC).** It refers to the area under the Receiver Operating Characteristic, a curve whose points defined in terms of Sensitivity and Specificity [86]. It is a measure of how far a model is from being a random guess [87]. Delving into this definition, the ROC curve shows the performance of classification model at all possible classification thresholds. This curve plots two parameters: the True Positive Rate (TPR) (i.e. the sensitivity) and the *False Positive Rate* (FPR), defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

The ROC curve plots TPR vs FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. Figure 2.14 shows a typical ROC curve.

The area under the ROC curve (AUC-ROC) provides an aggregate measure of performance across all possible classification thresholds.

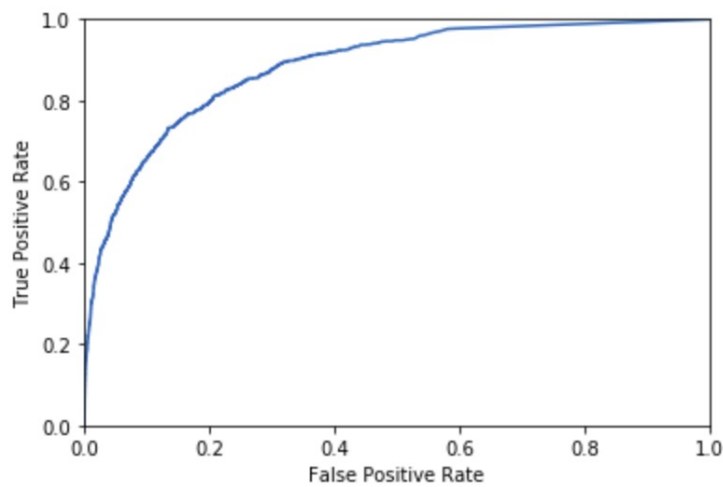


Figure 2.14: Example of ROC curve for a classification model. Every point forming the blue solid line represents a different (FPR, TPR) couple relative to a different classification threshold.

common way of interpreting AUC-ROC is the following: the probability that the model ranks a random positive example more highly than a random negative example. Accordingly, AUC-ROC ranges between 0 and 1. In particular, an AUC-ROC of 0.5 characterizes a random classifier while a perfect classifier has an AUC-ROC equal to 1. The worst classifier is characterized by AUC-ROC = 0.

Explaining algorithms' decision Shapley values

Machine Learning algorithms, correctly trained, can be used to accomplish a wide variety of important tasks, from credit card fraud detection [88] Alzheimer disease prevention [89], nevertheless what really prevents them from being more largely used is that they work as *black boxes* [90]: are fed into and an output is returned. The experimenter, except in some rare cases, does not directly check what features have mostly influenced their modeling its output [91]. However, controlling the path the algorithms follow in forming their decisions is important in understanding why they succeed and why they do not, separating confounding features from significant ones. Moreover, this transparency makes algorithms more reliable for a ever more widespread use.

Among all possible tools used to make algorithms more transparent, based on *Shapley values* [92], becoming increasingly popular. The Shapley value is a concept used in *game theory* [93] that involves fairly distributing both gains and costs to several factors working in a coalition. Game theory is when two or more players or factors are involved in a strategy to achieve a desired outcome or payoff.

Essentially, the Shapley value is the average expected marginal contribution of one player after all possible combinations have been considered. Shapley value helps to determine a payoff for all of the players when each player might

have contributed more or less than the others.

In order to apply this concept to Machine Learning, the notions of *game*, *players*, *payout* and *gain* must be defined in this context. In particular, the *game* is the prediction task of a single sample data; the *players* are the features involved in the prediction; the *payout* is the output value and the *gain* is the difference among the model's prediction and the average of all the outputs (i.e. the prediction done by a naive model). Accordingly, the Shapley values determine how to fairly distribute the payout among the features. Mathematically, it is possible to find a formula for calculating Shapley values [13].

In fact, in game theory terms, considering a N -player game together with a *value function*, v , that takes a subset of the players and returns the real-valued payoff of the game if only those players participated, the contribution of player i in the game, $\phi_i(v)$, is defined as follows:

$$\phi_i(v) = \sum_{S \subseteq \{1, \dots, N\} / \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S)) \quad (2.17)$$

In practical terms, equation 2.17 computes a weighted average payoff gain that player i provides when included in all coalitions that exclude i .

The Shapley values can be computed without any knowledge of the model's functioning. This makes it a model-agnostic technique, and also facilitates direct comparison of Shapley values for input features across different model types.

In addition, because Shapley values can be computed for the classification of every sample, they provide the granularity of local explanations while simultaneously allowing for global extrapolation. In fact, analyzing which features have higher Shapley values across multiple samples can give insight into the model's global reasoning.

Chapter 3

Startups and consumers' reviews case-cases and why they are important

In this chapter a fundamental concept of Industry 4.0 will be introduced: the *startup ecosystem* (or *innovation ecosystem*). It is defined as the set of startups and funders together with their funding relationships [94]. It is the boosts of technological innovation and consequently of countries' economic growth [95]. Accordingly, studying this system in depth will help investors in targeting their investments to the most promising firms in the most suitable countries. Furthermore, this task will be beneficial for startups too: they can identify the characteristics of the most successful ones (e.g., target market) and the corresponding investment [96]. There is an increasing interest in developing quantitative frameworks to identify and rate the strategic players of this ecosystem [97]. This task is particularly difficult to accomplish because of the dynamic and high-risk environment of startup companies [98, 99]. The first section of this chapter will deal with the datasets used in this work to analyze the startup ecosystem at different levels of detail.

As underlined in Chapter 1, another characterization of the Fourth Industrial Revolution is the *consumer-centric* conversion of firms. Accordingly, it is of paramount importance for firms to be reactive to the changing tastes and trends in order to be successful [100, 101]. As a consequence, the quantitative study of consumers' reviews and opinions to intercept feedbacks on products and facts is one of the main tasks that firms must carry out [102]. Therefore, in the second section it will be described a dataset containing tourists' reviews about their experiences in the accommodation facilities in Apulia, a region in the South-East of Italy. This purely data-driven approach will be effective in correctly highlighting the strengths of the Apulian tourist offer as well as the aspects to be improved. Since tourism is one of the most profitable business activities, especially in Italy, this study could have positive effects on both local and national economy. In the next chapter, the quantitative study of these data will be carried on using graph theory and Machine Learning.

3.1 The startup ecosystem's importance and modelling

In this section, first it will be clearly underlined why the startup ecosystem is important in the Industry 4.0 context, then, two of the most widely used datasets for the quantitative study of startups will be presented.

3.1.1 The importance of the startup ecosystem for Industry 4.0

The interplay between firms and investors is of paramount importance in determining the direction of the economic growth of a country [94]. Although there is an ongoing controversy about how growth and innovation are related, there is no doubt about the existence of a connection [95]. Accordingly, there is an increasing interest in developing quantitative frameworks to identify and rate the strategic players of economic systems. These studies aim at identifying the most promising and disruptive elements. The most evident example of the relations between startups and innovation is the *fintech world* [2]. Banks have increasingly crossed their paths with those of startups, through alliances, partnerships and incorporation so generating a progressive change of mentality in the sector.

Both small and large corporations are converting to the Industry 4.0 paradigm with a growing trend of industrial automation that aims at integrating new technologies creating new business models, increasing productivity and products quality. Firms' efforts aim at the creation of the *smart factory*, that is defined by three fundamental elements:

- **Smart production.** The new production technologies must promote collaboration between human operators, machines and tools.
- **Smart services.** The technological pervasiveness characterizing Industry 4.0 allows integration between systems, companies (e.g., customer relations), third party objects (e.g., roads, hubs, waste management facilities, etc.) and firms' customers.
- **Smart energy.** The energy consumption reduction is accomplished by data analysis. It aims at creating more performing systems and reduce energy waste according to the typical paradigms of sustainable energy.

The critical point, before being methodological, is about technology.

The contribution of startups plays a key role in pushing the existing technological limits beyond. In fact, one disruptive technology is the result of many tries and fails, and young startups are more prone to risk and experimentation. The impact that Industry 4.0-related startups are bringing will amount almost 200 billion euros in the next 2 years, and it will triple by 2030 [3]. Accordingly, characterizing the most promising startups is one of the main keys of success for both startups and investors.

As a consequence, studies aiming at quantitatively pointing out the most up-and-coming startups are gaining more and more ground. In fact, recently some scholars have introduced the concept of *high-impact entrepreneurship* [4, 5]. Startups' complex ecosystem, including investors, business angels, banks

and financing agents can help the understanding of the state of health of the economy and highlight the most promising and strategic firms. It is right to have just one possible definition of success within such an intricate system of relationships? Are there many possible and complementary definitions?

The analysis of complex economic ecosystems belongs to a research area, known as *science of success*, that is currently gaining considerable relevance [104, 105]. This emerging sector of complex system analysis takes advantage of the increasing availability of data to determine those patterns that underlie success in diverse areas, such as international country rankings [106], scientific publications [107], grant proposals [108], sport competitions [109] and patents [110]. The science of success investigates the impact of certain relations such as partnership, mentoring, collaboration or innovation, on the success of different initiatives, with the aim of identifying common good practices that could be applied in different contexts. The consequent results are often summarized through rankings: the entities on top are those employing the best practices while those in the lower part have room for improvement. Accordingly, in the next section, one of the most important rankings about the world countries' startup ecosystem will be presented.

3.1.2 Countries' innovation ecosystems: the StartupBlink ranking

Rankings are widely employed to quantify different kinds of performance. Their application range and importance are increasing, in the context of economics and politics [111, 112] and in private business [113]. Rankings significantly affect the process of decision-making, and their influence on the reputation of private and public institutions is extensively proven in literature [114]. A particular socio-economic aspect is beginning to be surveyed through rankings: the propensity and ability of an administrative region to create innovation ecosystems [115].

A necessary condition for an innovation ecosystem to raise and develop is the availability of economic resources and capabilities. In fact, they allow to create both products and business models with the required growth potential [116] and establish, at the same time, a community able to support the cooperation and interaction with investors [117].

In particular, entrepreneurs, investors and public policy-makers require as much information as possible about the available human and economic resources in order to assess a firm's possibility to survive and develop. Such an information request is satisfied both by rankings that measure economic systems performances, like StartupBlink [118] and Startup Genome [119], and more specialized databases like Crunchbase [120]. While Crunchbase structure and information content will be described in the next section, this one deals with countries' startup ecosystem rankings.

StartupBlink and Startup Genome have been introduced in 2016–2017 and published every year since then. They were the first worldwide rankings of innovation ecosystems. Nowadays, they are receiving increasing attention and diffusion in official press and social networks. In particular, the annual outcomes and rankings generate much interest in the startup community and among investors, as well as in government agencies, which often highlight their country's success in the international media emphasizing improvements in these rankings [121].

Nonetheless, the scientific community has paid little attention to innovation ecosystems and the analysis of their complexity. This work introduces a framework for the quantitative investigation of the multiple structural factors that condition their relevance and efficiency. In particular, this work will deal with the StartupBlink ranking, which provides information at a country level [122].

Even if rankings and their indicators provide an over-simplified representation of the complexity underlying cultural and economic phenomena, they nonetheless constitute one of the few quantitative tools used to explore the multifaceted aspects of social systems. Therefore, the use of rankings and indicators to set up government policies requires great attention to avoid critical issues. First of all, aggregate indexes may be influenced by arbitrariness and inaccuracy in choosing and aggregating different indicators, could even be partially correlated to each other [124, 125]. Second, interpreting a ranking could lead to ambiguities, since it provides a status-quo snapshot, that does not consider the heterogeneous starting conditions of the context in which a result is achieved. These differences are generally emphasized by rankings, while country performance assessment should be driven by the idea of similarity [126].

Detailed information on the development status can be useful for analysts and decision-makers to assess the result obtained by a given country in a ranking, because they allow comparisons with countries recognized as similar. Advantages of this approach are (1) it provides an equity-oriented criterion for the evaluation of a country performance, (2) it captures similarities among states that are essential for identifying and promoting possible unexpressed potentialities [96, 127].

The method proposed in this work relies on representing countries' development status in a complex and multifaceted way, replacing individual proxies determined by the arbitrary aggregation of indicators. This task will be accomplished in the next chapter by adopting the machinery of graph theory [39], which allows to represent and characterize interactions among constituents of a system. In particular, the graph will be built using the World Development Indicators (WDIs) database, a compilation of relevant, high quality, and internationally comparable statistics about global development and the fight against poverty [128]. This database collects yearly indicators starting, in the best case, from 1960, for 217 countries' economies (mostly belonging to the United Nations) and more than 40 economic or geographical country groups.

A crucial step of this analysis will consist of identifying network communities [47], namely non-overlapping groups of nodes with a tendency to create stronger connections inside the group than with the rest of the network (see section 2.2.1). The procedure defines a method to partition the set of countries based on their similarity, evaluated considering WDIs, and paves the way for a formulation of equity-based evaluation criteria. In fact, community detection actually keeps track of relevant similarities that in some cases can be hidden, unexpected and not deduced from merely geographical and economic considerations.

StartupBlink data and World Development Indicators (WDI)

Publicly available rankings about innovation ecosystems are an important and fairly recent tool. StartupBlink, in particular, was one of the first rankings to

be issued in 2016, and provides, nowadays, the most influential overview about the innovation ecosystems in the world [129]. It ranks the startup ecosystems of 100 world countries according to three main indicators:

- **Quantity.** It is determined by the number of startups in a country, presence of working spaces/accelerators (privately or publicly funded entities setting cohort-based program including mentorship [130]), startup events (pitch events in which startup founders present their ideas [131]).
- **Quality.** It is related to the impact of startups on their ecosystems. StartupBlink uses a variety of indicators to assess this index: startups' customer base, number of monthly visits on websites and number of Unicorns (private startup companies whose value exceeds 1 billion USD).
- **Business Environment.** It measures, based on the World Bank Doing Business report [132], the ease of doing business in a given country, considering aspects like the presence of technological infrastructures and the quality of bureaucracy.

The StartupBlink ranking considered in this work refers to 2019, pre-pandemic period, in order to avoid biasing effects on the ranking due to economic downturns triggered by the recent situation. The 2019 StartupBlink ranking, together with its component indexes, is reported in the Appendix (Table A.1). For simplicity, the countries listed in this ranking will be henceforth referred to as the *StartupBlink countries*.

As regards the WDIs, they will be considered only for the 100 StartupBlink countries. The choice of basing the network model on WDIs is due to the need for a development representation as multidimensional as possible. The WDIs database includes a wide variety of data but the indicators that will be used in this work are taken from the following categories: Environment, Economic Policy and Debt, Education, Financial Sector, Gender, Health, Infrastructure, Private Sector and Trade, Social Protection and Labor. These categories, in fact, cover essentially all the aspects of the development of a country.

The bulk file that we used for this study was updated to 15th September 2021. The dataset records 1443 WDIs, but missing entries are present in a variable number depending on the country. Data availability also changes with time, increasing, due to collection process improvements, from 1960 to the 2005–2016 period (a maximum is reached in 2010), and dropping in the following years, because some recent indicators are still unrecorded.

The choice to focus on 2019 indicators, motivated by the need to avoid pandemic biases, was also dictated by a tradeoff between recentness and data availability. In fact, missing entries in 2019 have been replaced from 2018 data or, in case of unavailability of the latter, from the 2017 dataset.

Moreover, indicators have been further selected following the criteria of data availability, consistency and information non-redundancy. This selection resulted in 426 indicators found applying the following sequence of actions:

1. Indicators with more than 10% missing values have been excluded.
2. To mitigate the effect of outliers, indicator values exceeding the 99th percentile and below the 1st percentile have been replaced by the reference percentiles.

3. Each indicator was scaled in the interval [0, 1] in such a way that 0 corresponds to the minimum value and 1 to the maximum.
4. To avoid redundancy, the Pearson correlation coefficient between all couples of indicators have been calculated. Then, the ones having a correlation value larger than 0.98 have been identified. Finally, for each of these couples the indicator having the smaller number of missing entries have been selected while the other has been excluded.

3.1.3 Interplay among startups and investors

A large body of literature defines a startup *success* according to its capability of obtaining massive capital [133]. Accordingly following the ideas of the *science of success* and *network success theory* [134], can be investigated the relation between the success of a startup and its ability in exploiting its own business network.

In particular, using a large public database, *Crunchbase*, the approach proposed in this work explicitly addresses the open questions raised by previous studies [135], especially concerning the possibility of success being strictly linked to a firm's networking. Few studies have investigated the economic systems of startup firms within quantitative frameworks [99, 105]. In this work, a quantitative framework for the analysis of interactions among startups and investors, together with a set of measures borrowed by graph theory. These metrics will allow to determine which elements in the startup ecosystem can be considered as *strategic* and which startups can be regarded as *successful* in the future.

Crunchbase contains large amount of data on the startup ecosystem, a special focus on investors, incubators, key-people, funds, funding rounds, and events. Crunchbase was created in 2007 by the TechCrunch company, managed it until 2015, when the Crunchbase platform became a private entity. According to OECD (Organisation for Economic Co-operation and Development), these data has been used for over 90 scientific publications [136], whose subjects range from business administration [137] to psychological evaluations of entrepreneurship [139], administrative science [140]. Particular mention must be given to studies concerning mathematical models, especially inspired by graph theory approaches [141, 89].

One question to be addressed is whether economic interplay can be accurately modelled with graphs, thus providing a quantitative and objective framework to define strategic and successful actors within an economic system. First of all, it will be demonstrated that the informative content extracted through classical statistical analysis fails to capture the whole picture. It must be noted that the information given only by funds (the only quantitative information of Crunchbase) fails to fully identify actors playing key-roles in the startup ecosystem. As a matter of fact, funds give no information about the number of the investors involved in a funding round. Moreover, the funds collected by a firm do not indicate its role in the setting. It does not yield any information about which firms are connected to it, or if it is a strategic element in the money conveyance.

Graph theory is an extremely efficient tool to model Complex Systems, especially to quantitatively highlight the importance of particular elements (i.e. nodes).

Accordingly in the next chapter it will be shown how the use of graph metrics points out the different roles played by economic actors and rank their importance. Moreover, it will be also shown how graph metrics can be also seen as a proxy of future success of startups. A startup is defined *successful* in a given year if it is an outlier of the distribution of the globally collected funds by all startups in that year. Then, these successful startups can be denoted also as *funding outliers*. In particular, a Supervised Machine Learning model will be presented and studied that relates the graph metrics of a firm to its possibility of being a funding outlier in a future time,

Crunchbase data

The Crunchbase dataset is formed by data collected on the *crunchbase.com* site. Specifically, the results presented in this work are based on its 13 October 2017 update. This site is, to date, widely considered as one of the most comprehensive publicly available dataset about investments and funding in the startup ecosystem on a global scale, as it contains more than 50 million records. More precisely, Crunchbase includes detailed information on more than 550,000 companies from 160 countries distributed among 38 different economic categories. Nonetheless, it is worth emphasizing that not all the companies and investors are involved in funding rounds, but 95% of them actually are. These latter elements are of interest for the subsequent analysis.

Some of these firms are investors, classified in 10 possible categories. Crunchbase data are organized in 17 distinct datasets, listed in Table C.1 in the Appendix, and focusing on several specific subjects, such as acquisitions, economic categories, collected funds, personnel, investment partners and geographic site, just to mention a few. Besides, Crunchbase includes different information about funding events (also denoted as *funding rounds*), how many funders are involved, how much money (in USD) was collected in a funding round and its date. In particular, funds are reported back to 1960.

Accordingly, it is possible to accurately track the flow and direction of investments and identify those companies that outperformed in attracting and/or investing capital.

Crunchbase companies are almost ubiquitous, nevertheless the USA is by far the leading country (53.6%). This is not surprising, being the USA an extremely favorable country for this kind of business; worth noting that the second country is the UK with only the 7.6%. Among different economic categories present in Crunchbase, Internet services and e-Payments are the most present, accounting for 19.3% and 14.4% respectively. Software (6.1%), science (5.8%) and ICT (5.6%) firms have also a non negligible representation. Finally, concerning the investor types, the most frequent ones are business angels (60%) and venture capitalists (28%), while other categories have occurrences not exceeding the 5% (see Table C.2 in the Appendix).

3.2 Why studying tourists' tastes in Industry 4.0

As underlined in Chapter 1, one of the most characterising aspect of the Fourth Industrial Revolution is the more central role of consumers in shaping firms' products and brand image [6]. As a consequence, a deeply understanding and

forecasting consumers' tastes and needs can be considered as the keys to firms' success [142].

Tourism is one of the most profitable economic activities worldwide and plays an important role in the economy of countries [143]. The advent and pervasiveness of new technologies in tourism started to transform radically the introduction of smart technologies (e.g smartphones) in the field of tourism has provided great possibilities for all its stakeholders (tourists, hotels, restaurants). In fact, not only tourists are now completely autonomous in booking facilities and means of transport, but, above all, they can give their opinions on services and experiences. This usually happens posting reviews on dedicated social networking services, like TripAdvisor. These reviews have the possibility to influence decisions and tastes of people worldwide. Since about 1.3 billion people travel around the world annually [144], all change in this area will truly have a big impact on the whole society.

Accordingly, unlocking the innovative potential of the entire tourism industry of a country by a deep insight of tourists' tastes and needs is of paramount importance for enhancing tourists' experiences and, above all, having a positive spillover on a country's economic status [18].

In particular, in this work it will be considered a database containing tourists' reviews about accommodation facilities in Apulia region in the South-East of Italy that has a strong and ever increasing tourist vocation [8]. This database will be described in detail in the next section.

3.2.1 Tourists' experiences in Apulia from TripAdvisor: reviews and rating

The importance of reviews in consumer decision-making has been widely confirmed by both academic research and practice [145]. The analysis of reviews has been based on different aspects such as volume, variation, perceived usefulness [145, 146] as well as their outcomes like review-based product rankings, trust in online reviews and management responses to consumer reviews [147, 148].

Research in tourism has highlighted online reviews as the major driver of brand choice [21], hotel performance [22], hotel bookings [23] and destination choice [24]. In particular, their effect on guests' satisfaction [25] has opened the discussion about numerical and textual aspects of the tourists' experience [26].

Numeric characteristics like the number of stars and the number of words included in a text have been studied in both decision-making [150] and customer satisfaction research [151]. However, the scalar ratings do not provide any information on those characteristics that customers like or do not like, while textual reviews display consumers' preferences, which can be extracted and analysed with specific techniques such as opinion mining and *sentiment analysis* [152].

Previous research [153, 154] used a mixed-method approach to analyze the numeric (ratings) and review text of online reviews to provide a deeper understanding of such a complex phenomenon. Recent studies [152, 155] have investigated the possibility to design and implement accurate systems to analyze the reviews and based on textual information predict their ratings. The variety of sources, languages and the different evaluating systems call for intelligent systems to use both textual and numerical reviews to better understand

the evaluation of the tourist experience and obtain useful information to improve the offer. The volume, subjectivity, and heterogeneity of social web-data require the adoption of specific methods combining Natural Language Processing (NLP) techniques to tokenize customers' reviews and carry out a subsequent sentiment analysis [156, 157].

However, an aspect that is often overlooked is that the reliability of these approaches is strongly affected by that of the ratings, misleading data, i.e. reviews with positive evaluations and negative ratings or vice versa, common due to psychological mechanisms such as social pressure [158].

The work carried on in this thesis aims to provide a unified framework to analyze the evaluation of the tourist experience and outline its key factors based on both ratings and textual reviews. In particular, this framework combines: data collected from TripAdvisor online platform (described later); sentiment analysis to detect the anomalous reviews whose score does not match with the measured sentiment; Machine Learning to train the classifier.

To explain how the considered models reached a decision and, therefore, to understand which factors were driving the tourist experience, the explainability framework based on Shapley values has been used, as explained in section 2.2.2.

The theoretical contribution of this study is twofold: (1) it contributes to the literature on online reviews by highlighting the impact of the combination of numbers and texts to help understanding and predicting tourist preferences; (2) it is one of the first studies using a cross-validation framework of the forecast model to avoid biased results based on the particular train-test subdivision of the dataset [159].

Moreover, sentiment analysis and classical Machine Learning methods are used in a fairly simple combination obtaining results comparable to those achieved with Deep Learning models [160], even though in a binarized-class problem, as explained in the next section.

The results obtained offer insights for practitioners and policy makers on how reviews should be analyzed to understand better their customers and improve their experiences.

TripAdvisor data

The tourists' reviews used in this work are collected using a web scraper. In particular, in September 2020 this tool was used to retrieve and download the reviews posted on TripAdvisor regarding the hospitality infrastructures in Apulia. Specifically, this dataset contains a total of 13399 reviews concerning 974 facilities, posted between May 2004 and June 2020.

TripAdvisor has been chosen for three main reasons: (1) it is one of the most accessed tourism-dedicated platforms, gaining more than 860 millions reviews and 8.7 millions opinions posted by more than five million registered users who visit the platform 30 million times per month on average [161]; (2) it considers heterogeneous facilities and tourism services, including accommodations, restaurants, airlines and cruises; (3) it includes a numerically-based rating system, through which developing a supervised model able to determine the rating from the corresponding textual information.

For each review six data fields are included:

- **rating.** The numerical score from 1 (bad experience) to 5 (excellent ex-

perience) that each user gave to the tourist experience.

- **review-id.** A 9-digit numerical code that identifies the review unambiguously.
- **struc-id.** A 40-digit alpha-numerical code indicating the accommodation facility.
- **struc-name.** The name of the facility identified by the *struc-id*.
- **date.** The date the review was input in the platform.
- **vicinity.** The address of the reviewed facility.

Only reviews in English have been considered because they reduce the potential bias of language. Moreover, NLP tools for the pre-processing of English texts are well consolidated with respect to other languages [162].

In terms of rating the data are highly unbalanced; more than half of reviews represents an excellent experience (numerical score equal to 5), 27% are given a score equal to 4, 10% are related to a score equal to 3, while less than 10% reviews have a numerical rating of 2 or less.

In other terms, considering as positive those reviews having a score greater or equal to 3 [146, 21, 163], the number of positive reviews is much greater than the negative ones. This imbalance is commonly observed in studies dealing with services' reviews [18].

The first step in transforming textual data before feeding them to Machine Learning algorithms is *tokenization*, every text element (words and punctuation) is considered as element of its own called *tokens*, punctuation and *stop-words* are removed. Stop-words are those words that are useful in building texts but are meaningless (articles, conjunctions, prepositions). After that, the remaining words are *lower-cased* in order to avoid repetition of words differing just for lower and upper casing letters. Finally, the words are *stemmed*, in order to obtain their root [164]. Consistently with previous studies [146, 21] ratings have been *binarized*, reviews with rating lower than 3 were considered *negative* and labeled as 0, while those having a rating higher or equal to 3 were considered *positive* and labeled as 1.

Then, a sentiment analysis of the reviews has been carried out and the users' ratings have been compared with the measured sentiment. It is common to have a mismatch between the review's rating and its sentiment [165], reviews are defined as *contradictory* if those belonging to class 0 (negative-rated) have a positive sentiment and vice-versa. Therefore, 1460 reviews are filtered out, about 9% of the sample. The cleaning step is important for the framework's reliability despite the fact that these reviews represent 9% of the whole dataset, as explained in the next chapter.

Since the dataset is highly imbalanced (less than 10% of reviews are negative), positive reviews have been *undersampled* in order to obtain unbiased classification models, that is, a number of positive reviews equal to that of negative ones are randomly chosen. Accordingly, a perfectly balanced dataset have been obtained, containing all the negative reviews and a subsample of positive reviews. Then, this balanced dataset have been fed into machine learning algorithm. This approach have been repeated 100 times.

Chapter 4

How to boost innovation and customers' satisfaction: deploying graph theory and Machine Learning

The startup ecosystem has been presented in Chapter 3 together with its importance in the context of the Fourth Industrial Revolution. In particular, two datasets have been introduced and described: the *StartupBlink* ranking that rates world countries' effectiveness in creating an innovation ecosystem (section 3.1.2) and *Crunchbase*, that contains information about the economic interplay between startups and investors (section 3.1.3). In the first two sections of this chapter, it will be shown how these systems can be modelled using graphs, how useful information can be extracted using network theory algorithms and how Machine Learning methods can be used to discover hidden patterns and make predictions. Together with the startup ecosystem, Chapter 3 highlights the importance of understanding tourists' tastes and needs through the quantitative analysis of their textual reviews. In particular, a dataset of tourists' reviews extracted from TripAdvisor has been introduced (section 3.2) dealing with accommodation facilities in Apulia (a region in the South-East Italy). These reviews describe their experiences and from users' point of view, the characteristics of the Apulian tourism. The third section of this chapter will show how Machine Learning combined with Natural Language Processing (NLP) techniques and Explainability tools will be able to highlight, in a purely data-driven way, those aspects that should be improved in the Apulian tourism offer and those that represent an asset to focus on.

4.1 StartupBlink: equity oriented rethinking through community detection

In this section it will be first shown the graph model used to study the StartupBlink ranking. Then, the community detection algorithms used to unveil

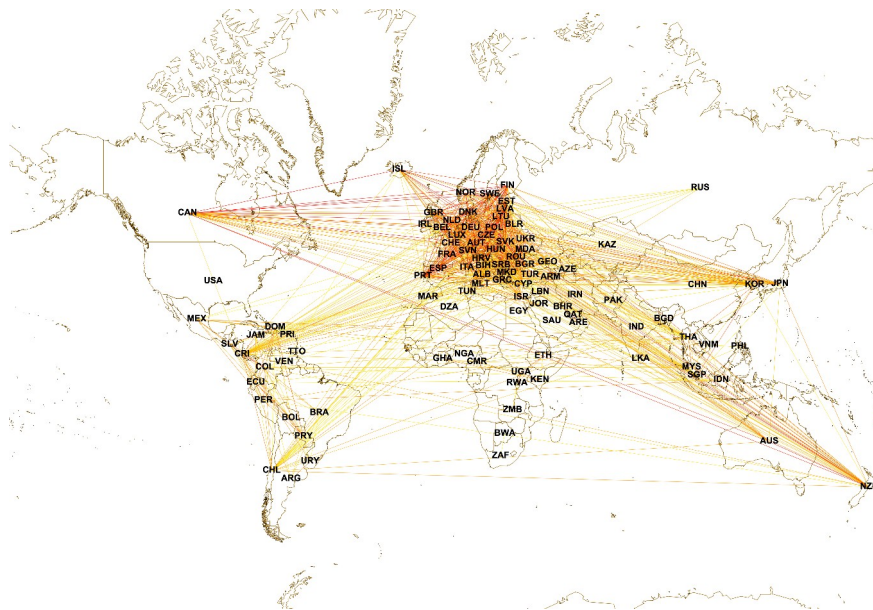


Figure 4.1: Part of the network of StartupBlink countries based on the values of their WDIs. Only edges whose weight is equal or greater than 0.70 are reported. Link colors are related to their weight, in ascending order from yellow to red. Source: [122].

similar groups of countries will be exposed together with other mathematical tools needed to accomplish the task of the equity oriented rethinking of StartupBlink.

4.1.1 StartupBlink country network

As shown in section 3.1.2, the StartupBlink countries have been characterized by 426 World Development Indicators (WDIs). Then, these WDIs have been employed to evaluate the corresponding pairwise Pearson correlations. A graph has been built as follows: each StartupBlink country is represented by a node; pairs of nodes are connected by weighted edges whose weight is determined by the pairwise Pearson correlation between the sets of WDIs associated to the corresponding countries. In particular, only those links whose Pearson correlation is statistically significant (at 1% significance level) are retained. Thus, a connected network of 100 nodes with 482 weighted links is obtained. A geographical distributed version of the network is depicted in figure 4.1.

4.1.2 Community detection algorithms and Resolution Ratio

In this section the algorithms used to perform community detection on the StartupBlink country network will be first presented. Then, a novel mathematical tool is introduced, the *Resolution Ratio*, that proves to be helpful in relating the community membership of a country with its performance in the StartupBlink

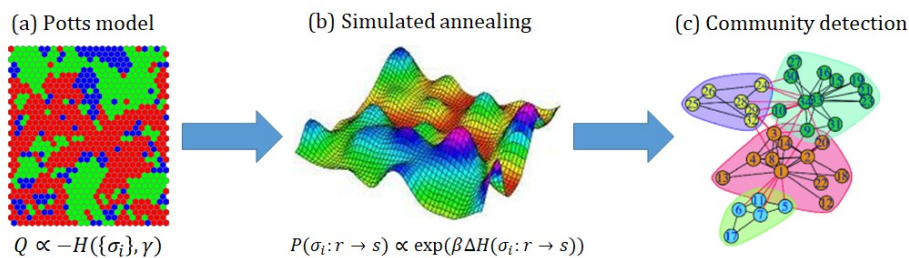


Figure 4.2: Flowchart of the Spin Glass algorithm. (a) A Potts model is built from the graph. A spin state, representing a community index, is assigned to every node. An Hamiltonian is defined that is proportional to the modularity and that depends on a coefficient, γ , representing the role of non-existing links among communities. (b) The next step is finding the ground state of the system. It is found using the simulated annealing heuristic method based on assigning a transition probability to every node. This probability depends on the difference of H between transitions and on the *cooling factor*, β , that represents the *temperature* of the spin system. (c) Once the ground state is reached, the corresponding spin states of the nodes represent the community labels.

ranking. The Resolution Ratio is the fundamental step for the equity oriented rethinking of StartupBlink.

Spin Glass and Leiden community detection algorithms

Since the Pearson correlation can be both positive and negative, community detection algorithms must take into account this characteristic of the graph. Therefore, algorithms that are suitable to handle signed weights must be used [47]. The most commonly used are *Spin Glass* [167, 50] and *Leiden* [168]. *Spin Glass* uses the simulated annealing technique to optimize modularity. *Leiden* is a hierarchical clustering algorithm that recursively merges communities into single nodes by greedily optimizing the modularity and the process repeats in the condensed graph (see section 2.2.1).

In particular, the Spin Glass algorithm workflow is depicted in figure 4.2. It consists of the following steps:

- **(a).** A Potts model [169] is built from the graph. The community index of the i -th node $\sigma_i \in 1, \dots, c$ is interpreted as a spin value with c possible values. An Hamiltonian is built taking into account the weighted links between nodes (having both positive and negative values) [50], that represent their interaction. The Hamiltonian depends on a parameter, γ , that weighs the contribution of non-existing links among communities. It can be demonstrated that the following formula holds [167]:

$$Q(\{\sigma_i\}) = -\frac{1}{m}H(\{\sigma_i\}) \quad (4.1)$$

where m is the number of links in the graph.

- **(b).** According to equation 4.1, modularity maximization is equal to the minimization of the Hamiltonian. Then, the next step is to find the

ground state of H . This task can be accomplished using heuristic methods, since it is a NP-hard problem [50]. Accordingly, the simulated annealing algorithm is used. This technique is based on swapping the spin values (i.e. the community indexes) of nodes and considering the corresponding change in the Hamiltonian. For example, if σ_i changes its value from r to s , then the corresponding change in the Hamiltonian is denoted as $\Delta H(\sigma_i : r \rightarrow s)$. In particular, a swapping probability is assigned to every node, depending on $\Delta H(\sigma_i : r \rightarrow s)$ and another parameter, β , or *cooling factor*, that represents the *temperature* of the Potts model: $\frac{1}{T}$. This probability has the following form:

$$P(\sigma_i : r \rightarrow s) = \frac{\exp(\beta \Delta H(\sigma_i : r \rightarrow s))}{\sum_{j=1}^c \exp(\beta \Delta H(\sigma_i : r \rightarrow j))} \quad (4.2)$$

The goal of this step is to find the distribution of community indexes that, according to the swapping probability of equation 4.2, minimizes H .

- **(c).** Once the ground state has been found, the corresponding distribution of spin states (i.e. community indexes) among the nodes represents the set of communities found by the algorithm.

As regards the Leiden algorithm, it is a refinement of the Louvain algorithm [168]. In fact, it is well acknowledged that the latter can return badly connected communities [169]. Leiden workflow is depicted in figure 4.3 and the corresponding steps are the following:

- **(a).** Initially, every node represents a community on its own.
- **(b).** Every node is merged to other communities (with other nodes) in order to find the best partition one that maximizes the modularity. In the Leiden algorithm, modularity has the following form:

$$Q = \frac{1}{2m} \sum_{ij} A_{ij} - \gamma \frac{k_i k_j}{2m} \delta(c_i c_j) \quad (4.3)$$

where an additional parameter, γ , is present. This parameter is called *resolution* since it determines level of detail of the communities to be found: higher resolutions lead to more communities, while lower resolutions lead to fewer communities.

- **(c).** A refinement of the partition with maximum modularity is found. In fact, every community is treated as a graph on its own and subdivided into smaller communities, following a modularity maximization approach, in (a). This refinement marks the difference with the Louvain algorithm. In the refinement phase, nodes are not necessarily greedily merged with the community that yields the largest increase in the modularity, a node may be merged with any community for which the quality function increases. The community with which a node is merged is selected randomly [170]. The larger the increase in the quality function, the more likely a community is to be selected. The degree of randomness in the selection of a community is determined by a parameter γ .

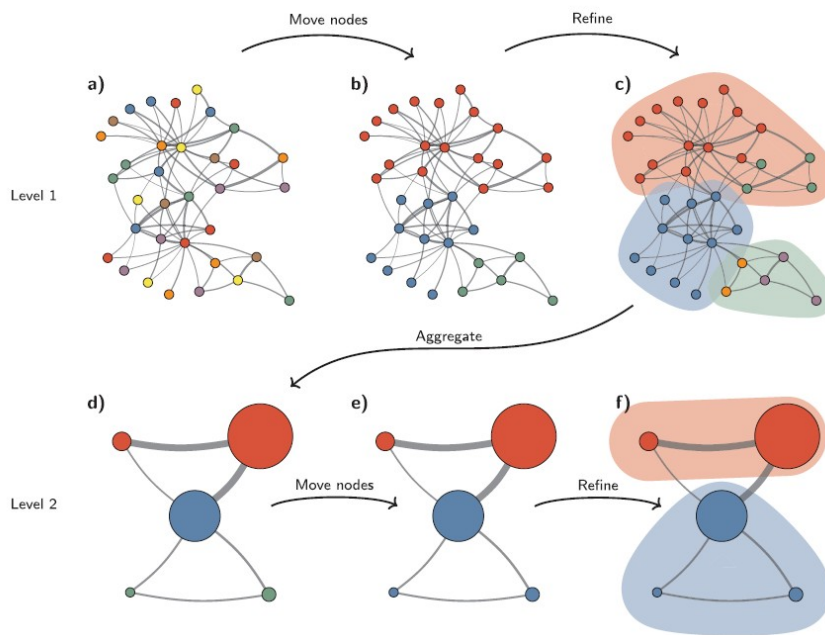


Figure 4.3: Flowchart of the Leiden algorithm (a) Every node is initially in its own community (b) Nodes are merged to form communities that maximize modularity. This merging happens randomly with the only constraint that modularity must be increased by this merging (c) These communities are refined. (d) The nodes in the communities are aggregated (e) The aggregated nodes are moved in order to maximize modularity (f) Another refinement process is carried out. Source: [168].

in the selection of a community allows the partition space to be explored more broadly. Node mergers that cause the quality function to decrease are not considered.

- **(d).** An aggregate network is created based on the refined partition, using the non-refined partition to create an initial partition for the aggregate network. For example, the red community in (b) in figure 4.3 is refined into two subcommunities in (c), which after aggregation become two separate nodes in (d), both belonging to the same community.
- **(e).** Then, the individual nodes in the aggregate network are moved as in (b).
- **(f).** The refinement procedure described in (c) is applied even in this case. These steps are repeated until no further improvements can be made.

Moreover, for both the considered algorithms, a hierarchical community detection by recursive partitioning has been performed [171, 172]. This procedure employs a multi-step process in which the detection algorithm is applied subsequently in order to find a subdivision of communities coming from the previous stage.

This procedure stops when an iteration condition is no longer satisfied. This condition is determined by the accordance between outputs of different runs of the algorithm. It should be noted that both community detection algorithms are not deterministic, thus providing different outputs when applied to the same graph. Nonetheless when community detection is robust, the outcome should be as independent as from randomness. Moreover, the output of community detection also depends on the choice of the Spin Glass or Leiden algorithm parameters.

Accordingly, in order to choose the right parameters for the community detection algorithms and obtain consistent communities, a criterion has been used: one of the algorithms is used to partition the network 100 times; same outcome occurs in at least 90% cases, that partition is accepted, and recursive partitioning proceeds to the next step, otherwise the iteration stops, and the partition found at the previous level is accepted as a final result.

This method is used with both community detection algorithms together with an accurate exploration of their parameters space. In particular, the Spin Glass algorithm as explained before depends on two parameters: the resolution γ and the cooling factor. γ ranges in the interval [0.5, 1.5] with a step of 0.1; the cooling factor has values in the interval [0.1, 0.9] with a step of 0.1, besides the extreme values 0.01 and 0.99.

The Leiden algorithm depends on the resolution γ as well and the randomness, β . The resolution varies in the same range as for Spin Glass, while β ranges in [0.01, 1] with a 0.01 step, besides the extreme 0.001.

The choice of parameters is determined by the request of output consistency and robustness with respect to their variations.

The performance of the community detection algorithms is analyzed upon varying parameters, by monitoring the behaviour of three quantities:

- percentage of agreement. This parameter is computed for a given set of parameters as the ratio between the number of occurrences of the most common network partition and the total number of runs of the algorithm.

- Number of communities in the most common partition.
- the *inverse participation ratio* (IPR) in the most common partition defined, for a partition in K subsets of a network with N nodes, as follows:

$$IPR = \frac{1}{\sum_{i=1}^K \frac{n_i}{N}} \quad (4.4)$$

where (n_1, n_2, \dots, n_K) are the cardinalities of each subset. The IPR is a coefficient used to evaluate the number of communities among which the considered network can be considered *effectively* shared. For example a partition in $K = 3$ communities of a network of $N = 90$ nodes is characterized by $IPR = 3$ if $n_1 = n_2 = n_3 = n_4 = 30$, while a partition with cardinalities $n_1 = 60; n_2 = 20; n_3 = 10$ yields $IPR = 1.976$, much closer to 2 than to 3.

The overall pipeline of community detection has been implemented in Python 3.8 using the *networkx* library and the built-in functions for both Spin Glass and Leiden algorithms. The corresponding code has run on a single machine endowed with a Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz and a Windows 10 Home operating system. The entire community detection pipeline took 1 hour to be executed.

Reinterpreting the ranking Resolution Ratio

In order to quantify the connection between community membership of a country and its performance in the StartupBlink ranking, the *Resolution Ratio*, R , has been used [106, 22]. It is defined as follows: consider a partition of N elements, to which the values $1, \dots, N$ are assigned, in K disjoint groups with cardinalities n_c with $c = 1, \dots, K$.

One can associate to the full distribution of x an overall mean value μ and a variance σ^2 . On the other hand, given the partition in groups $c = 1, \dots, K$, one can evaluate for each group the related mean μ_c and variance σ_c^2 .

The definition of R is based on the fact that the overall variance can be viewed as composed of two positive contributions [173]:

$$\sigma^2 = \sigma_{int}^2 + \sigma_{ext}^2$$

$$\text{where } \sigma_{int}^2 = \sum_{c=1}^K \frac{n_c}{N} \sigma_c^2 \text{ and } \sigma_{ext}^2 = \sum_{c=1}^K \frac{n_c}{N} (\mu_c - \mu)^2$$

Considering that σ_{int}^2 is the weighted average (with weight n_c/N) of group variances, whereas σ_{ext}^2 is determined by the discrepancy between group means and the full distribution mean, the quantity

$$R = \frac{\sigma_{ext}^2}{\sigma_{int}^2} \quad (4.5)$$

is an indicator of how much group distributions tend to separate.

In the considered case, groups coincide with network communities and when the distributions of a StartupBlink score (i.e. the global StartupBlink score or the Quantity Quality and Business index) corresponding to different communities have small overlap with each other, the resolution ratio tends to be much larger than 1, while it becomes very small if community distributions fully overlap.

$R \approx 1$ can be considered as an intermediate case with mean values of neighboring community distributions separated by an amount that is close to the typical inter-community variation of the considered network. Therefore, $R = 1$ can be assumed as a threshold value that separates cases in which reading country performances in the light of communities is either meaningful or not.

In the next section the main results will be presented.

4.1.3 WDI country communities and StartupBlink rethinking

After outlining the methods to be applied to obtain an equity oriented rethinking of the StartupBlink ranking, this section presents the relevant findings.

First, the results of partitioning the network of StartupBlink countries in communities will be presented. Then, the obtained subdivision is quantitatively compared with the well-established income-based country groupings employed by the World Bank. Finally, the performances of countries in the StartupBlink rankings are reinterpreted (for both the global StartupBlink index and its three constituent indexes: Quantity, Quality and Business) based on their community membership, provided the distribution of the corresponding index in communities satisfies $R > 1$.

WDI country communities

As described in section 4.1.2, two different algorithms have been used (Spin Glass and Leiden) exploring a wide range of the related parameter spaces, obtain a hierarchical community detection framework.

The robust partition of the StartupBlink countries graph found through this process consists of three communities, will be labelled henceforth as (I,II,III). The geographical distribution of countries in these communities is shown in figure 4.4

Both Spin Glass and Leiden algorithms stop after two iterations of the hierarchical pipeline described in section 4.1.2, providing the same split in each step. In particular, in the first iteration the algorithms return two communities, comprising 49 and 51 countries, respectively.

Then, in the second iteration the first community splits in two sets, composed of 22 and 27 countries, while the second community, composed of 51 nodes, cannot be subdivided anymore. Therefore, the final partition of the StartupBlink graph consists of three communities. The membership of countries to these three final communities is reported below, with countries identified according to their ISO-3166 alpha-3 code standard [174].

- *Community I (22 countries):* USA, GBR, CAN, ISR, AUS, NLD, SWE, CHE, DEU, FRA, FIN, IRL, DNK, SGP, JPN, BEL, NZL, AUT, NOR, LUX, ISL, MLT;
- *Community II (27 countries):* ESP, EST, RUS, LTU, KOR, POL, CZE, ITA, CHN, PRT, CHL, UKR, BGR, SRB, ROU, HUN, GRC, LVA, SVN, SVK, HRV, BLR, MKD, MDA, CYP, PRI, BIH;
- *Community III (51 countries):* IND, MEX, THA, COL, BRA, ARE, IDN, TUR, ARG, MYS, ZAF, KEN, PHL, NGA, PER, EGY, PAK, GEO,

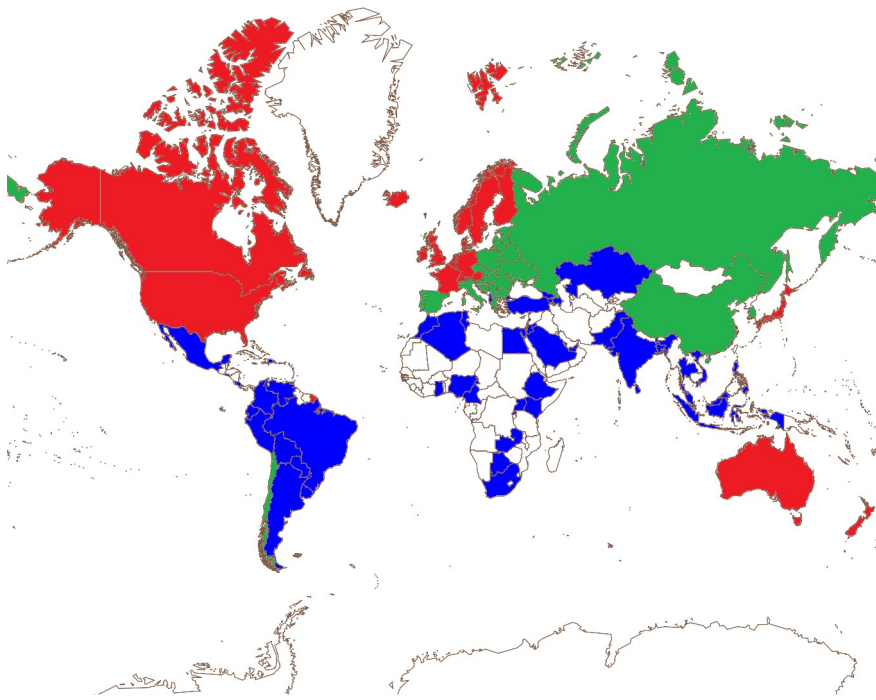


Figure 4.4: Communities of StartupBlink countries, determined from the similarity graph based on WDI indicators. Community I (red) contains 22 countries; community II (green) contains 27 countries; community III (blue) contains 51 countries. Source: [122].

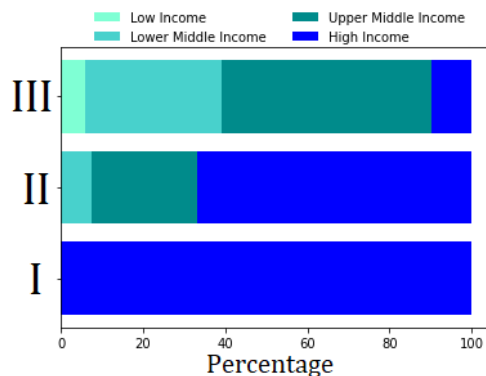


Figure 4.5 Composition of StartupBlink network communities in terms of World Bank income groups, highlighting an emerging hierarchy in descending order from I to III, in terms of income source [122].

ARM, RWA, MAR, AZE, KAZ, URY, VNM, JOR, TUN, GHA, ECU, LKA, DOM, SAU, UGA, LBN, IRN, CMR, ALB, CRI, BGD, JAM, BWA, SLV, ZMB, VEN, TTO, BHR, PRY, QAT, BOL, DZA, ETH.

Interestingly, as can be observed from Figure 4.4, many states that are members of the same community have also geographical boundaries in common, together with the economic one. Comparison with the partitions determined by the the World Bank income groups [175] indicates reported in Figure 4.10, that communities are ordered in a descending way from I to III in terms of income: therefore, the expression *wealth communities* will be used henceforth when referring to them. Even though this result is not surprising from an economic point of view, it has been found in a data-driven and unsupervised way. This proves that the complex network approach developed in this work is effective in representing the real economic situation and can be used as a quantitative basis to extract useful insights.

Rethinking StartupBlink ranking in the framework of wealth communities

The partition in communities represents both a way to group countries based on socio-economic similarities but above all a mean to reinterpret their outcome in the StartupBlink ranking.

In fact, it is reasonable to expect a tendency of ranking index values referred to the same community (i.e. GlobalStartupBlink index and Quantity and Business index) to cluster together and separate from the values related to other communities. Accordingly, one could point out both those countries whose performances in the ranking go beyond the expectations determined by community membership and the ones that, on the other hand, underperform with respect to their community performance. However, such an assumption can be considered valid only after being checked *a posteriori*. Figure 4.6 represents, through violin plots, the distribution of all the StartupBlink indexes. The vertical coordinates corresponding to the considered index values, while the horizontal coordinate is determined by country community membership.

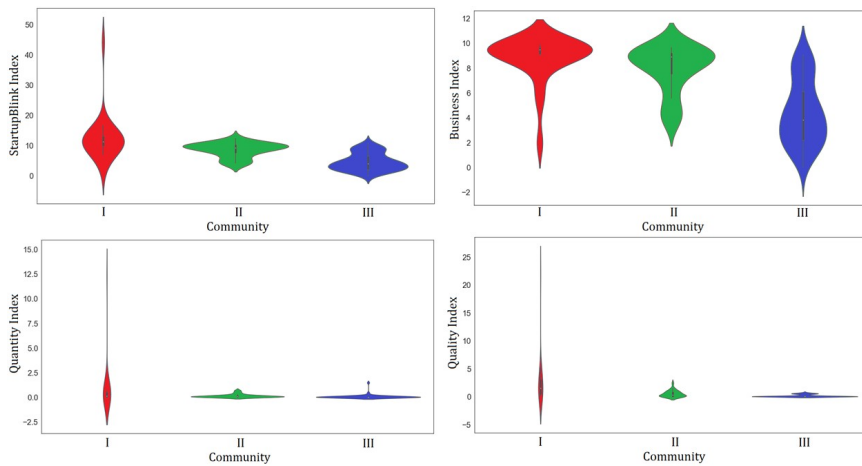


Figure 4.6: Violin plots of the distributions in the three wealth communities of the overall StartupBlink index (top left panel) and of the specific Business (top right), Quantity (bottom left) and Quality (bottom right) indexes. It can be observed that the plots related to the overall and the Business index show a tendency of community distributions to separate from each other. This tendency, will be confirmed by the analysis of the resolution ratio [122].

StartupBlink index	Quantity index	Quantity Index	Business Index
2.661	0.863	0.148	1.789

Table 4.1: Resolution ratio values for the global StartupBlink index and its constituents in bold, R values greater than 1. Source: [122].

From figure 4.6 it can be observed that the three communities show clearly different distributions for both the global StartupBlink and the Business index. On the other hand the Quantity and Quality indexes have the greatest part of their values centered around 0 in the three communities. Community I is endowed with few higher values for both these indexes, represented by the longer upper tail of the corresponding violin plots. Even though these observations must be confirmed quantitatively by the Resolution Ratio, it can be expected that the Quantity and Quality indexes are not able to characterize the communities. This would indicate that the country startups are settled do not significantly impact their development in terms of the attributes described by the Quantity and Quality indexes (the presence of coworking spaces, accelerators and startup events, startups' customer base).

According to section 4.1.2, the Resolution Ratio is used to quantify how much country performances in the considered rankings are related to the corresponding community members. Considering the global StartupBlink index together with its three components, the corresponding Resolution Ratios values are reported in Table 4.1.

Resolution Ratios relative to the StartupBlink global index and the Business index are both above 1. This means that wealth communities are well resolved with respect to both the index measuring the easiness of business in a country (Business index) and the indicator quantifying the global value of its

innovation ecosystem (StartupBlink index).

Since $R > 1$ for two indexes, reasonable community-based predictions can be made on country performances in the rankings defined by these two indicators. Moreover, the performances diverting from the expected outcome can be critically evaluated. In particular, *top-of-the-class* countries in a given ranking are defined as those whose score falls, at the same time

- beyond the 75-th percentile of the community they belong;
- beyond the 25-th percentile of at least one higher-wealth community.

An analogous criterion is applied to define *room-for-improvement* countries, as those whose score is placed satisfies these two conditions:

- under the 25-th percentile of the community they belong;
- under the 75-th percentile of at least one lower-wealth community.

Top-of-the-class countries should be taken as models by those states that are similar in terms of development and aiming at improving their status in the considered ranking.

The mismatch of countries' performances and the community-based expectation can be further characterized by assigning a symbol for each 25-th percentile of a higher-wealth community that is overcome by it. On the other hand, *room-for-improvement* countries are the ones that have the potential of achieving better results in the ranking, reaching those of countries in similar development conditions. In this case, a further characterization of performance can be provided by marking a country with a symbol each time the score lies under the 75th percentile of one lower-wealth community.

Countries having the highest scores in community I or those with the lowest scores in III do not fit the previous definitions, since it is not possible to compare their results with more or less developed communities, respectively.

Thus two specific categories have been introduced to classify these remarkable performances. *Benchmark countries* are those belonging to community I and characterized by a score beyond the 75th percentile of community. They can be viewed by the rest of the world as best-practice examples. On the contrary, *trailing countries* are those belonging to community III, with their scores smaller than the 25th community percentile. These states could require ad-hoc support to improve both their political and economic practices and improve their innovation ecosystem. Below is reported the complete evaluation of country performances as measured by StartupBlink index and Business index, in accordance with the aforementioned criteria:

StartupBlink index

- **Community I.** *Benchmark:* USA, GBR, CAN, ISR, AUS, NLD; *Room-for-improvement* : NZL (*), AUT (*), NOR (*), LUX (*), ISL (*), MLT (*).
- **Community II.** *Top-of-the-class:* ESP (↑), EST (↑), RUS (↑), LTU (↑), KOR (↑), POL (↑), CZE (↑); *Room-for-improvement:* BLR (*), MKD (*), MDA (*), CYP (*), PRI (*), BIH (*).

- **Community III.** *Top-of-the-class:* IND (↑), MEX (↑), THA (↑), COL (↑), BRA (↑), ARE (↑), IDN (↑), TUR (↑), ARG (↑), MYS (↑); *Trailing:* BGD, JAM, BWA, SLV, ZMB, VEN, TTO, BHR, PRY, QAT, BOL, DZA, ETH.

Business index

- **Community I.** *Benchmark:* USA, GBR, SWE, FIN, DNK, NZL; *Room-for-improvement :* ISR (*), BEL(*), NOR(*), LUX(*), ISL(*), MLT(*).
- **Community II.** *Top-of-the-class:* ESP (↑), EST (↑), LTU (↑), KOR (↑), POL (↑), CZE (↑), PRT (↑); *Room-for-improvement:* MKD (*), MDA (*), CYP (*), PRI (*), BIH (*).
- **Community III.** *Top-of-the-class:* IND (↑), MEX (↑), THA (↑), COL (↑), BRA (↑), ARE (↑), IDN (↑), TUR (↑), ARG (↑), MYS (↑); *Trailing:* BGD, JAM, BWA, SLV, ZMB, VEN, TTO, BHR, PRY, QAT, BOL, DZA, ETH.

4.2 Crunchbasegraph model and forecasting success

In this section it will be first presented the graph model of Crunchbase data and the network metrics used to extract information from this network, after checking that this information cannot be retrieved using classical statistical analysis. A Supervised Machine Learning model will be presented aiming at identifying the startups that will be successful in the future using network metrics. As underlined in section 3.1a, a startup is denoted as *successful* in a given year if it is an *outlier* in the distribution of funds that are collected from all the startups in that year.

4.2.1 Modelling the economic interplay

The economic interplay shown by the Crunchbase dataset can be naturally modelled with a directed complex network. Nodes represent all the elements reported in the dataset, both startups and funders, and the directed links correspond to the investments. In particular, the source node is the investor while the target is the element receiving funds. The reason for such a model is twofold: on one hand, this representation is adherent to traditional economic approaches monitoring the money flux; on the other hand, this model of economic interplay is straightforward and easy to interpret.

Moreover, thanks to this model, quantitative assessment of node importance can be provided. Thus, it can be established to which extent a firm plays a strategic role within the economic system and establish how the success probability of its business is related to its network properties.

Denoting with N the number of Crunchbase economic players involved in fundings and L as the set of the registered economic transactions (or *funding rounds*), for each pair of nodes $n_i, n_j \in N$, a transaction $(n_i, n_j) \in L$ represents a flux of money from n_i to n_j . Accordingly, the directed graph G denoted as the couple (N, L) , has order $|N| = 121950$ and size $|L| = 289396$. This

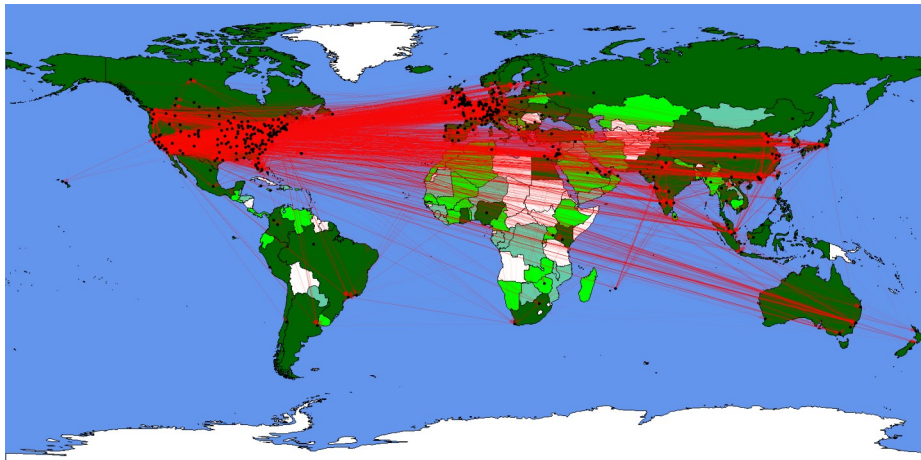


Figure 4.7: A pictorial world map of the Crunchbase ecosystem nodes are in black and arrows in red. For each nation the percentage of companies tracked within Crunchbase is represented through different shades of green. The darker the shade, the greater the number of nodes in that country. Source: [105].

graph is not symmetric as the existence of a connection does not imply the existence of its counterpart ($n_j n_i$). It is worth noting that the network model is built using all transactions occurred between 1960 and October 2017. Figure 4.7 depicts a snapshot of the network model of Crunchbase.

It should be noted that Crunchbase does not keep track of the amount of each transactions, so that a weighted graph model is not possible. Nevertheless, the overall amount of collected funds for each company is known. Considering the amount of collected funds f as the variable representing the business success of each company and given the country c , the economic category e and the investor type t as auxiliary attributes, each node can be parametrized as: $n_i(f; c, e, t)$.

Even if f is a fundamental measure of nodal importance, in the next chapter it will be first demonstrated that it does not yield a thorough picture of startup ecosystem. On the contrary, the network properties can quantitatively assess the flux of capitals and can significantly improve its description.

As explained in section 2.1, there are different network metrics representing different flavours of importance within the network [177]. The measures expressed by network centralities represent information complementary to that provided by collected funds and help highlighting different points of view on the startup ecosystem. For example, *degree* centrality measures the overall number of connections of a node. The larger the number of connections the greater the importance of the node. Another example is the *betweenness* centrality. This measure evaluates the importance of a node by taking into account the number of paths within the network passing through that specific node. Accordingly, because of the directed network model for Crunchbase, three centrality metrics have been considered for characterizing each node: degree (equation 2.3), outdegree (equation 2.2) and betweenness (equation 2.11).

Accordingly, the firms are considered *strategic* if their behaviour in terms of funds, degree or betweenness significantly differ from other firms.

These three measures have been chosen as they have a direct economic interpretation:

- **Indegree** represents the node's attractiveness to investors.
- **Outdegree** accounts for node's financing power;
- **Betweenness** represents the node's capacity of capital conveyance.

Being able to take account of this information implies a deeper knowledge on the economic system of startup firms; it takes into account not only the funds, but even how they are collected, sourced and conveyed. It is worth noting that, in general, these distinct actions can be performed by different agents. Besides from this picture it is also manifest that considering only the amount of funds collected by a firm provides too limited a description of the system.

4.2.2 Defining and measuring success

As underlined previously, straightforward definition of *success* for a startup business is the amount of capital it is able to collect. This definition is robust both in terms of meaningfulness and interpretability. Moreover, capitals are measurable, thus providing an objective strategy to evaluate success. Of course, success is a multifaceted concept and can be defined in many different ways, e.g. by considering the startup acquisition as a successful result. However, these aspects represent complementary viewpoints with their own peculiarities and interpretation difficulties. In characterizing the startup system conclusion, the choice to consider success as firms according to collected funds is twofold: (i) it is intuitive and (ii) widespread in economic literature [133, 134].

Even though the amount of funds collected by a startup can be considered a reliable measure of success, it provides a limited picture. For example, the amount of collected funds does not contain information about the number of investors and does not quantify the attitude to convey capitals within the system.

For example, within an economic system there are firms whose main role is not that of collecting capitals, but rather investing them. Accordingly, their importance would be hidden if only the amount of collected funds would be considered. Nevertheless, their presence is an invaluable asset for the functioning of the whole ecosystem. Another crucial aspect deals with the way capital moves throughout the startup ecosystem. In graph theory it is well known that some nodes can deeply influence other nodes even when they are not directly connected, but thanks to an indirect influence.

Comparing the distribution of all the collected funds with those of indegree, outdegree and betweenness a statistically significant difference has been found ($p < 10^{-3}$, through the non-parametric Kolmogorov Smirnov test) between all centrality distributions and the funding amount reported in the Appendix (figure C.1).

This analysis confirmed that the information conveyed by network centralities does not significantly overlap with that provided by *Trends*. For each distribution, the outlier observations have been determined. Since the distributions considered in this work are all positive definite, the strategic companies are

	Fundings	Indegree	Outdegree	Betweenness
Country	USA (70%)	USA (72%)	USA (57%)	USA (55%)
	CHN (7%)	UK (5%)	UK (7%)	CHN (9%)
	UK (4%)	CAN (3%)	DEU (3%)	IND (5%)
Economic Category	Is (21%)	Is (25%)	eP (72%)	eP (24%)
	Sc (15%)	Sc (9%)	Is (6%)	Is (21%)
	eP (8%)	eP (8%)	Sc (2%)	Sc (7%)
Investor type	VC (63%)	VC (54%)	VC (50%)	VC (51%)
	PE (15%)	Acc (35%)	Ang (31%)	Acc (25%)
	Inv (4%)	HF (4%)	PE (9%)	PE (10%)

Table 4.2 Best performers for Nationality, Category and Type comparison of top three rankings according to Fundings, indegree, outdegree and betweenness. Countries are abbreviated according to International Naming Convention. Categories: Internet services (Is), Payments (eP), Science (Sc) Investor Types: Venture Capital (VC), Private Equity (PE), Accelerators (Acc), Business Angels (Ang), Investment bank (Inv), Hedge Funds (HF). Source: [105].

precisely defined as the *right outliers* of the corresponding distributions. These elements are able to collect funds, investors, investments, and capital transfers significantly better than others.

A standard procedure to define the outliers employs the *boxplot* method. For each centrality measure, all the elements whose values exceed the threshold value given by the 75-th percentile of the corresponding distribution added to $1.5 \times$ the interquartile range (IQR) are defined to be an outlier (precisely a right outlier). In this sense they are *strategic* companies. Further methodological details are provided in Appendix C.5.

4.2.3 Strategic elements in the startup ecosystem

Following the procedure shown in the previous section, 176 outliers have been found for the distribution of funds; as regards the centrality metrics, 4716 outliers have been determined for indegree, 1284 for outdegree and 523 for betweenness. Besides the bare numeric differences, further insights have been obtained by considering the Kendall correlation, τ , between each centrality distribution and that of the funds. This coefficient measures the degree of monotone relationships between funds and network metrics in ranking the elements of the startup ecosystem. Results reveal that the indegree centrality has the highest correlation with funds, $\tau = 0.4$ at 1% statistical significance. On the other hand, outdegree and betweenness are less correlated ($\tau = 0.1$ for both of them at 1% statistical significance). The top 50 firms for each ranking are reported in the Appendix (Table C.3) whose synthetic overview is presented in Table 4.2.

These findings show that the ranking of Crunchbase elements according to funds has a negligible correlation with that obtained using outdegree and betweenness. This means that these two network metrics convey a different information from funds and, consequently, allow to extract insights not otherwise obtainable. As regards indegree, even if it has a stronger correlation with funds, which is an intuitive result, there is not a perfect correlation. This means that

having a high number of investors does not necessarily imply high funds, the latter can come in solitude. In other words Crunchbase depicts both crowd-funding situations and funding rounds in which a small number of operators invest high funds.

A further characterization can be provided instead in terms of economic categories and investor types. The sole inspection of funding outliers has unveiled important information about successful results on top nations, economic categories and investor types reported in Table 4.2. Do network metrics either confirm these findings or provide insights? Accordingly, the funding outliers have been compared with the indegree, outdegree and betweenness ones and significant differences have been found for nationalities, economic categories and investor types (using the Kendall tau coefficient, $p < 0.05$ Bonferroni corrected).

In particular, these analyses have underlined the role played by USA and Chinese firms for what concerns nationalities; Science and Internet services for economic category; finally, venture capital, private equity, accelerator and business angel for investor type (Fig. 4.8).

Further details about this analysis are presented in the Appendixes C.2-C.10. USA firms are able to collect more funds than expected just looking at network metrics, the larger difference being between funding and outdegree. This result is not a surprise since USA host the majority of Crunchbase firms and provide extremely advantageous economic conditions, especially for startups. It is instead surprising that the prevalence of USA firms among outdegree outliers is much smaller (around 20%) than for the other distributions. One of the reasons is the fact that USA firms are the most frequent among the Crunchbase elements importantly affects these results; nevertheless, the fact that a country is present with a given frequency does not ensure that its attributes (indegree, outdegree and betweenness) should be outliers with the same frequency. Fig. 4.8 shows that this happens only for USA and China. In these nations it can be observed a significant difference between the frequency of funding outliers and graph centralities. In particular, USA firms can collect more funds than expected just looking at network centralities, the larger difference being between funding and outdegree.

The startup ecosystem comprises almost entirely the set of possible economic sectors. Through the analysis of how funds are distributed among successful firms, it can be established that Science applications and Internet services are generally the economic categories that collect the largest amounts of funds. In fact, these two categories account together for about 38% of funding outliers. On the contrary, network centralities, especially outdegree and betweenness, outline the role played by e-Payment firms. Actually, e-Payment firms represent the 72% of outdegree outliers and the 23% of betweenness outliers, which makes sense as this specific economic sector is particularly devoted to capital investments and conveyance [180].

As regards investors, four significant outcomes can be highlighted: Venture Capital firms have a prevailing presence among outdegree outliers, according to their compelling vocation for investments. Private equities show a significant presence among outdegree and betweenness outliers. Contrary they are absent from indegree and funding outliers. This suggests that their strategic role in conveying investments. The important fraction of

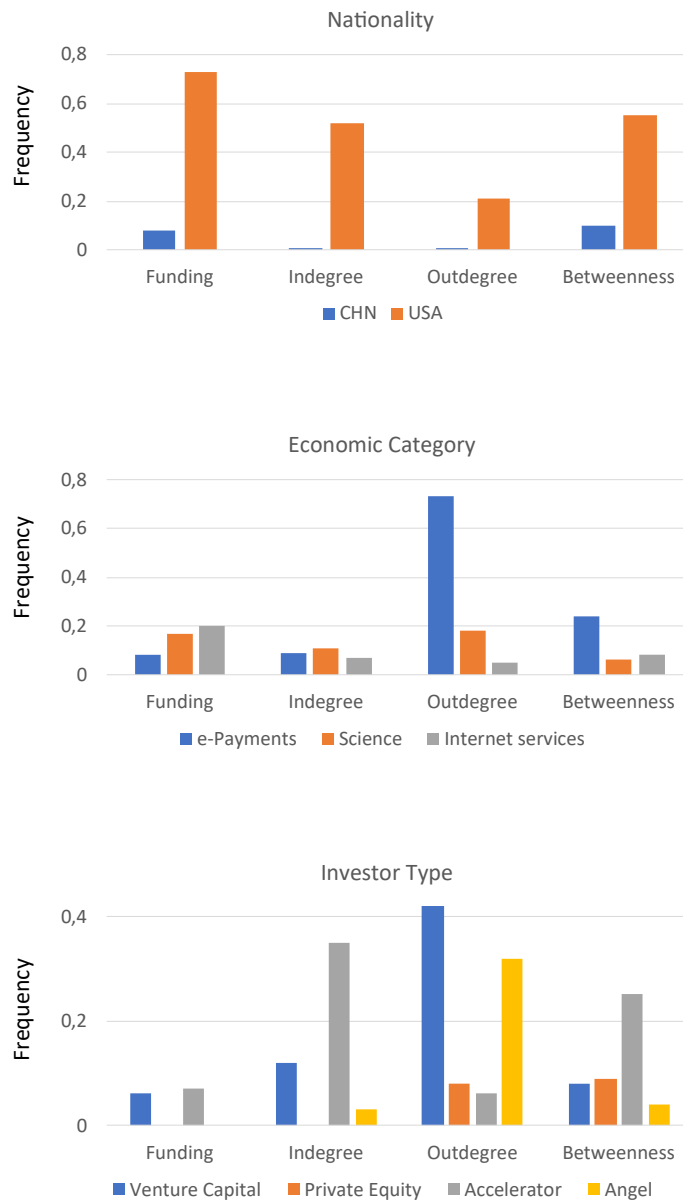


Figure 4.8: Top performers frequencies of nationalities (top), economic categories (middle) and investor types (bottom) in the outlier funding and centrality metrics distribution. Source: [105].

accelerators among indegree and betweenness outliers suggests an interesting interpretation: strategic accelerators are oriented to collect funds from a large number of investors and convey them to other firms, they are strategic players, acting as connection elements within the startup ecosystem. The special role in the startup ecosystem would have been neglected in an analysis based only on funding outliers. In fact, in the latter, accelerators represent only 7% of firms, while their frequencies among indegree and betweenness outliers are 38% and 24%, respectively. (iv) The outdegree outliers show a significantly larger presence of business angels (36%) compared with other distributions. The result depicts the fundamental role played by these investors in granting funds to a large number of firms. Even in this case, this role would not be noticed by just looking at the funding distribution, where business angels do not appear at all.

4.2.4 Forecasting success

In this section two fundamental questions are addressed: can we identify successful firms with the outliers of funding distribution, network centralities proxies of this notion of economic success? If yes, to which extent? Considering that in Crunchbase each firm is represented by a node enriched with attributes, $n = n(f; c, e, t)$ (see section 4.2.1), an alternative formulation is searched that models funding f by means of the corresponding network metrics $n = n(i, o, b)$ where i, o and b are the proposed centrality measures: degree, outdegree and betweenness, respectively.

It is worth noting that, based on the peculiar nature of the startup funding, a usually a one-time-event, the amount of collected funds in one funding round is weakly correlated to those raised in successive funding rounds, as demonstrated in figure 4.9

In fact, the figure shows how correlation is weak even at low values of future years. For example it is 0.2 at 1-future year and approaches zero as the time interval between the two observations increases.

From figure 4.9 it can be observed that the variance of correlation of fundings decreases with future years. This means that it may happen that, one or two years after receiving higher (lower) funds, a startup may need other higher (reslower) funds. Since it is a newly established business, it is a reasonable phenomenon. Moreover, since startups are usually subject to mentoring programs with enforced steps, they must demonstrate their business potential as quickly as possible. They must prove to be economically self-sufficient in the long-term otherwise they become unattractive for investors. The funds a startup receives four years after the first funding round are not related to those it has received in the past, but are linked to the business value it has developed in those years.

Even though multiple supervised Machine Learning algorithms could be applied, for the sake of interpretability and given the exiguous number of dependent predictor variables, a logistic regression model has been chosen (see section 2.2.2). The logistic regression has manifest advantages: it returns both a measure of importance for each predictor, given by the magnitude of coefficients, and the direction of association, namely the sign of coefficients. Nonetheless, other learning and modeling strategies (Random Forests, Deep Learning) could be adopted and could represent an interesting theme for future works.

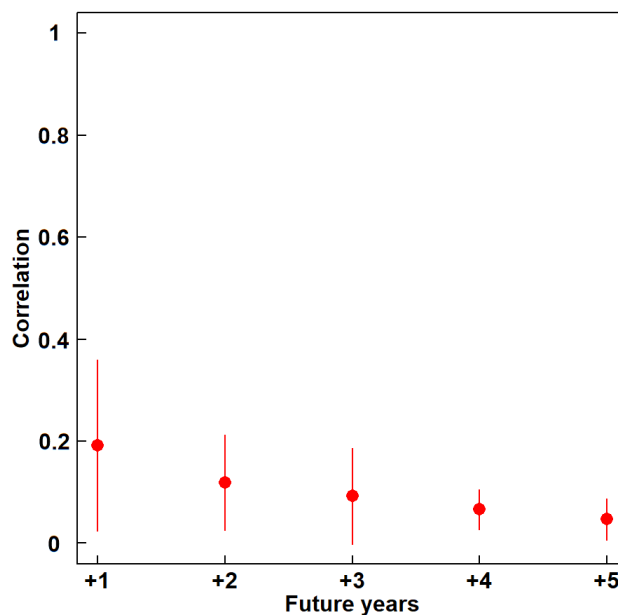


Figure 4.9 Mean correlation variation, along with its standard deviation, with future years Source:[105].

Formally, the outcome variable f is 1 for a successful firm and 0 otherwise. In formula:

$$f = \frac{e^{\beta_0 + \beta_1 i + \beta_2 o + \beta_3 b}}{1 + e^{\beta_0 + \beta_1 i + \beta_2 o + \beta_3 b}} \quad (4.6)$$

where β_s are the coefficients of the logistic regression that measure the impact of the network metrics in determining the probability of future success of a startup. β_0 is a *bias term* determining the success probability independently of the network metrics. These coefficients are determined in the training phase of the logistic regression algorithm.

Equation 4.6 returns a real value in the interval $[0, 1]$ so that a startup in the test set is denoted as *successful* if $f \geq 0.5$, otherwise it is classified as *not successful*.

Since until 1999 only 39 funding records are present, while they are 10 just considering the year 2000, only data referred to years ranging from 2000 to 2017 are considered for this task, thus resulting in 298 firms.

For each year T and for each node, the three network metrics have been considered: the in-degree (i), the out-degree (o) and the betweenness (b), which are the independent variables of the model. The dependent variable, $f(T)$ indicates whether node n in year T is an outlier for the corresponding distribution of collected funds or not.

In order to accomplish this task, for every year $T \in \{2000, \dots, 2017\}$ the corresponding network is built and the nodal centralities have been computed; then, for each node it is determined if a future year $T + 1, T + 2, \dots$ it corresponded or not to a funding outlier. Successful firms have been labeled with 1 and 0 otherwise.

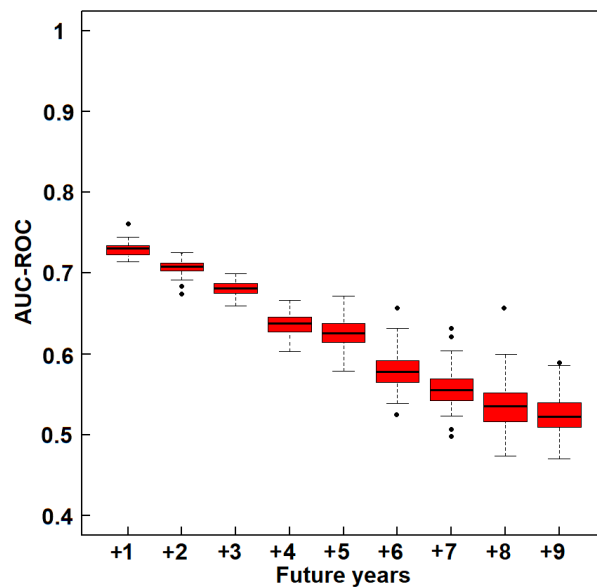


Figure 4.10 AUC-ROC of predictions up to nine years obtained with the logistic regression model source:[105].

Then, the logistic regression model have been trained at a time T while the future years have been used for test.

The analysis has been carried out within a 5-fold cross-validation framework and the procedure has been repeated for 100 iterations, this whole procedure is used to predict whether after 1, 2, . . . , 9 years a firm will be a funding outlier and evaluated the performance of the model in terms of the AUC-ROC, whose results are shown in figure 4.10

These results show the presence of a correlation between network centralities and the amount of collected funds up to four/five years in the future with median AUCs ranging from 0.73 (+1 year) to 0.61 (+5 years). As expected, the forecasting accuracy decreases with the increasing of the forecast time horizon: the prediction to 9 years is barely distinguishable from random.

Besides sensitivity and specificity have been analyzed together with their variation according to the ratio between successful and unsuccessful firms for each year, figure 4.11

Two considerations arise following these results: (1) the logistic model's ability of retrieving non-funding outliers (i.e. specificity) slightly grows over time; (2) the drop in the performance observed in terms of AUC-ROC values is caused by the worsening of sensitivity, i.e. the capability to detect successful firms. This effect is dominated by the substantial drop of these firms over time, in fact the successful firms which initially represent the 4/5% of the data, after 9 years are only the 1%

To evaluate the importance of the different predictors, Cohen's D [181] has been chosen. Cohen's D is an effect size measure; it compares the difference of two sets of observations or measures with their intrinsic variability:

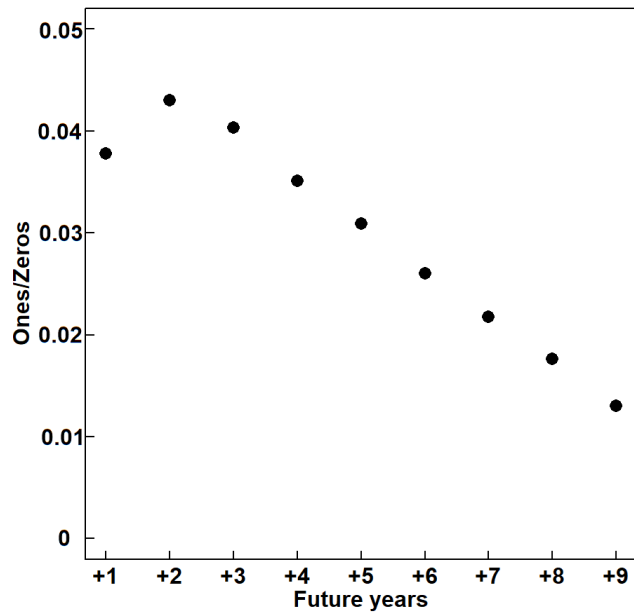
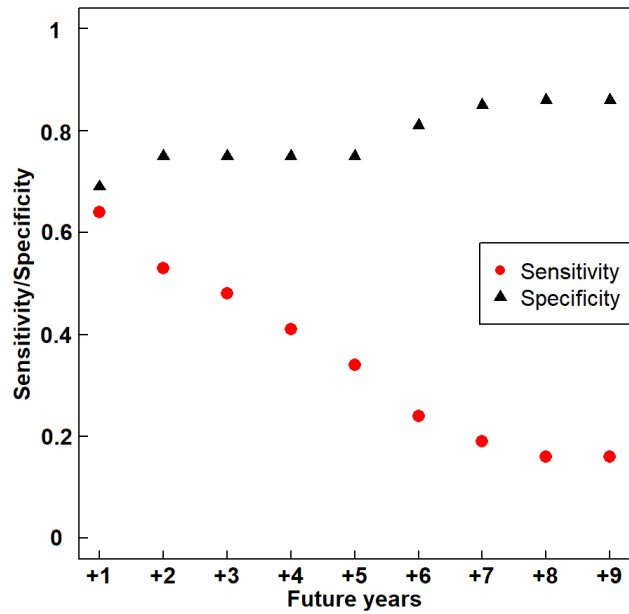


Figure 4.11: Sensitivity and specificity analysis. Top panel shows Sensitivity and specificity results of the model. Bottom panel: The ratio of successful elements (Ones) over the *not successful* ones (Zeros) as a function of the future years. It can be seen that as the years in the future grow the number of successful elements becomes much smaller than those of the non successful ones: the task of predicting success in the future becomes ever more difficult with the increasing of the forecast horizon. Source: [105].

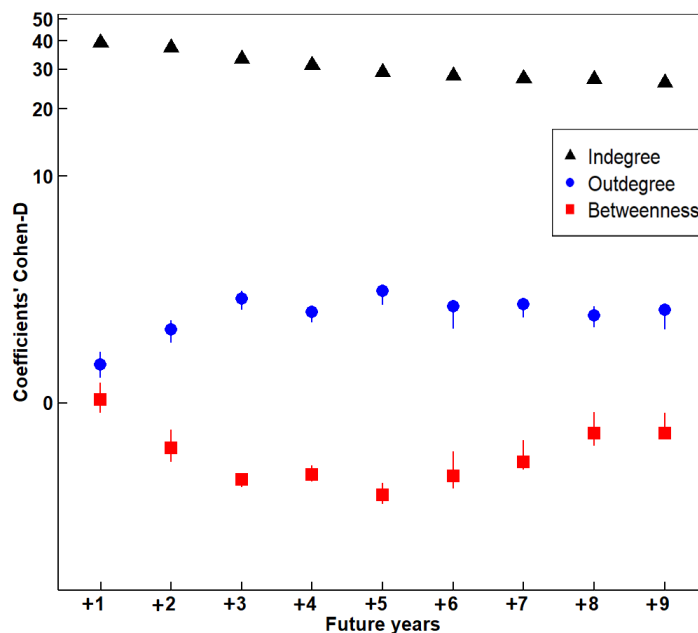


Figure 4.12: For each year the importance of node centralities is expressed in terms of Cohen's D coefficients in symmetric log scale. This scale emphasizes that, even though outdegree and betweenness have a smaller impact than indegree on the future success of a startup, indegree has a positive effect on startups' probability of future success, while betweenness has a negative impact. Source: [105].

$$D = \frac{E[X_1] - E[X_2]}{\sigma}$$

where $E[X_1]$ and $E[X_2]$ denote the expectation values for the sets of observations X_1 and X_2 , respectively; σ is the pooled standard deviation divided by the square root of the number of training observations. Using Cohen's D to evaluate the feature importance in the logistic regression, it can be found that the indegree is the most relevant feature to predict success in collecting funds, see figure 4.12

This result is particularly evident at very short time ranges (+1 year); interestingly, at time scales between +1 year and +3 years, the effects of both outdegree and betweenness increase. For larger times, the indegree still stands as the most important predictor, the other centralities remain comparable, but with different signs.

These results can be interpreted as follows: in the long period the successful firms are not only those able to collect capitals from many investors, but also those playing an active role in financing other firms;

Interestingly, the more an element is able to facilitate the money conveyance in the startup ecosystem, the more its probability of having success in future years decreases. This result indicates that even if money conveyance can be considered an asset [135], it should be considered with caution when collecting

funds.

However it should be taken into account that the startup-funding is the only funding mechanism considered here. Moreover it should be highlighted that many betweenness outliers are stable and powerful (e.g. Alibaba, Google, Yahoo, Amazon, Uber) which obviously do not focus their activities on collecting funds in the startup ecosystem, have a fundamental role as publicly acknowledged mentors, thus justifying their prominent role in conveying money.

4.3 TripAdvisor: extracting insights from tourists' reviews

In this section the process used to transform reviews' textual data into a suitable format to feed Machine Learning models will be first shown, the results of these algorithms will be presented and, finally, the explainability results will be discussed in order to point out strengths and weaknesses of the Apulian tourism offer.

4.3.1 From text to numbers TF-IDF matrix

After the textual processing analysis described in section 3.2.1 (tokenization, lower-casing and stemming) the main tool used to transform textual unstructured data in a mathematical form is the *Term Frequency - Inverse Document Frequency matrix* (TF-IDF matrix) [182, 183]. This matrix has each element defined as the product of two factors:

$$TF_{(i,j)} = \frac{n_{ij}}{|d_j|}; \quad (4.7)$$

$$IDF_{(i,j)} = \log \frac{|D|}{d(i)}, \quad (4.8)$$

where n_{ij} is the number of occurrences of word i in the j -th review, D is the set of all the reviews in the dataset, $d(i)$ is the number of reviews in D containing word i at least once, and $|S|$ denotes the cardinality of a set.

The term $TF_{(i,j)}$ rewards the frequency of a word within a review: the more cited word i in review d_j , the greater the importance of d_j . The term $IDF_{(i,j)}$, on the other hand, penalizes the ubiquity of a word in all the considered reviews and underlines the rarely occurring terms. In fact, a word that is widely used in all texts does not allow discrimination among them. The complete TF-IDF matrix has dimensions 848×1698 . 848 is the number of the considered reviews, while 1698 represents the length of the *vocabulary* of the reviews, i.e. the set of unique words derived from the textual processing analysis. Moreover, the corresponding binary rating is assigned to each review. This last variable is the output to be predicted through Machine Learning models fed by the words in the vocabulary representing the input features.

Then, these data have been fed to Machine Learning models and the respective performance have been evaluated using the metrics described in section 2.2.2, as shown in the following section.

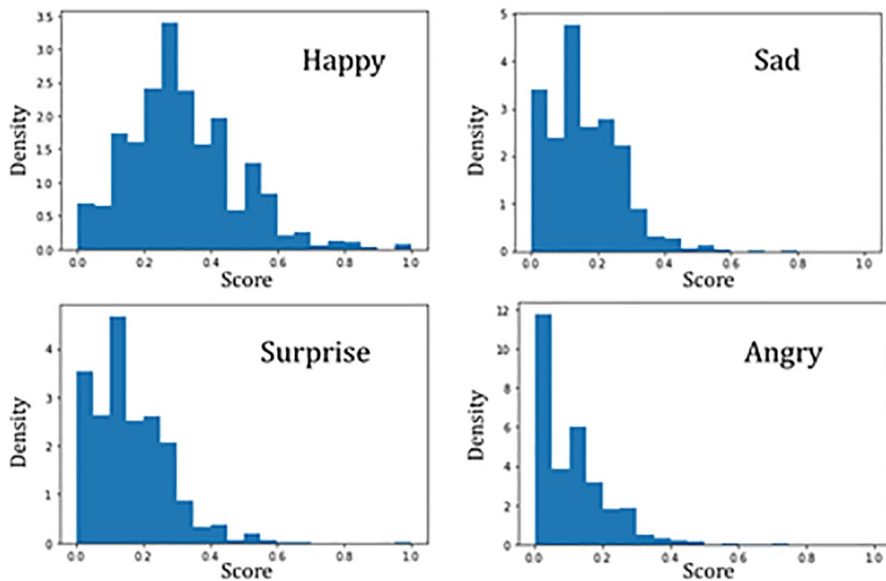


Figure 4.13: Score density distribution of the emotions. Each review was enriched with four continuous scores (one for every emotion) and the scores were normalized. Source: [83].

4.3.2 Reviews' classification

As underlined in section 3.2.1, reviews are asymmetrically distributed in terms of ratings. In particular, the positive reviews (those having a rating greater than 3) accounted for the 91% of the entire dataset. Since the rating distribution is highly skewed in favor of positive reviews and a different threshold for their binarization would have not yielded any significant differences, the undersampling technique is chosen to balance the data to avoid any bias in the learning models [184]. To set up an effective rating forecast model, the emotions expressed in the reviews have been studied, so that not only contradictory reviews have been highlighted (see section 3.2.1), but also the emotions that mostly affected the model's performance. This sentiment analysis has been performed using the VADER framework [185], one of the most widely used tools to accomplish the analysis of sentiments in social networks. In fact, this analysis highlighted four emotions: Happiness, Sadness, Anger, and Surprise. The intensity of the emotions showed that the reviews express happiness more than other emotions, thus confirming a positive experience, as shown in figure 4.13.

Figure 4.13 clearly shows that negative emotions (Sad and Angry) have their score distributions shifted towards low values. This phenomenon is well acknowledged in the literature and is unavoidable, as well as the overwhelming presence of positive reviews with respect to the negative ones [18]. Nonetheless, this phenomenon does not represent a bias in the following analysis, since the undersampling technique has been used in order to balance the presence of negative and positive reviews (see section 3.2).

Table 4.3 shows the classification performance of the classical Machine Learning models used: Random Forest (RF), Gaussian Naive Bayes (GNB),

Model	acc(%)	AUC(%)	F1(%)	sens(%)	spec(%)
RF	89 (82-95)	96 (91-99)	89 (82-95)	89 (82-95)	92 (82-98)
GNB	77 (70-85)	83 (76-85)	77 (68-82)	76 (69-85)	81 (68-91)
SVM	88 (82-91)	94 (91-96)	85 (80-88)	88 (80-86)	87 (76-88)
XGB	83 (76-91)	93 (86-97)	84 (77-90)	88 (76-91)	84 (73-94)

Table 4.3: Models' performance measures obtained by filtering out contradictory reviews. The metrics reported in the table are accuracy (acc), AUC-ROC (AUC), F1-score (F1), sensitivity (sens), specificity (spec). Metrics' values are reported as percentages and the values in parentheses are the 5th and 95th percentile, respectively. Source: [83].

Model	acc(%)	AUC(%)	F1(%)	sens(%)	spec(%)
RF	62 (58-66)	66 (62-70)	62 (58-66)	62 (58-66)	60 (54-65)
GNB	54 (49-60)	58 (53-62)	54 (60-60)	54 (49-59)	61 (58-64)
SVM	60 (57-64)	60 (58-65)	59 (55-62)	58 (54-63)	61 (58-65)
XGB	58 (55-62)	62 (58-64)	56 (53-61)	60 (57-64)	58 (55-63)

Table 4.4: Models' performance measures obtained including contradictory reviews. The metrics reported in the table are accuracy (acc), AUC-ROC (AUC), F1-score (F1), sensitivity (sens), specificity (spec). Metrics' values are reported as percentages and the values in parentheses are the 5th and 95th percentile, respectively. Source: [83].

Support Vector Machine (SVM, with linear kernel) and Extreme Gradient Boosting (XGB). These results are obtained in a 5-fold cross validation framework repeated for 100 times. In particular, table 4.3 shows the mean values of the performance metrics as obtained from cross-validation, while in parentheses are shown the corresponding 5-th and 95-th percentile.

In particular, RF model scores are significantly better than those of the other models in all the measured metrics. This has been established using a Mann-Whitney statistical test ($p < 0.01$). Nonetheless, all models reached satisfactory levels of accuracy. Also, these values are significantly enhanced with respect to the performance obtained using the whole dataset, reported in Table 4.4. This result is consistent with the literature on the impact of noisy data (i.e., the contradictory reviews) on Machine Learning algorithms [186, 187].

This result ensures that this measurement relies only on the informative content provided by the reviews and not from the specific algorithms adopted.

4.3.3 Strengths and weaknesses of the Apulian tourism offer

Finally, in order to highlight the key factors driving model classification, it must be calculated the input features contribution to the classification score. As explained in section 2.2.2, this can be done using the Shapley values. These quantities are shown in figure 4.14 for the first 20 words in decreasing order in terms of the absolute mean Shapley value.

The words with the highest importance are *breakfast*, *work* and *staff*, which

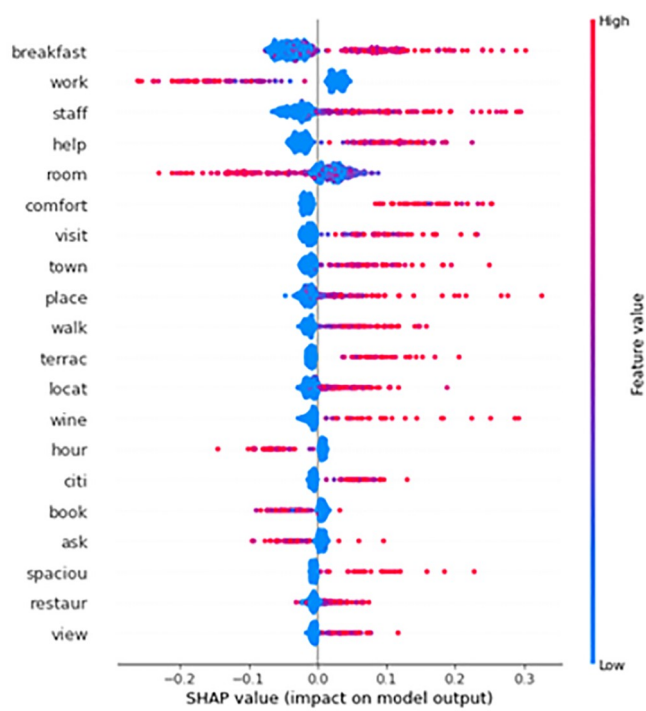


Figure 4.14: The Shapley values of the first twenty important words. The contributions towards a positive or a negative review are distinguished according to the frequency a word appears within the text (high/low) [83].

1	great	17	bad	33	welcom	49	meat
2	friendli	18	work	34	hotel	50	dinner
3	love	19	recommend	35	waiter	51	say
4	breakfast	20	best	36	book	52	charg
5	excel	21	tell	37	host	53	walk
6	town	22	area	38	enjoi	54	air
7	help	23	room	39	terracc	55	terribl
8	comfort	24	amaz	40	stai	56	leccc
9	perfect	25	water	41	fantast	56	atmosph
10	nice	26	delici	42	star	58	wine
11	staff	27	ask	43	disappoint	59	free
12	good	28	place	44	rude	60	view
13	poor	29	park	45	worst	61	hour
14	locat	30	clean	46	highli	62	fresh
15	beauti	31	relax	47	definit	63	local
16	wonder	32	pool	48	arriv	64	peopl

Figure 4.15: The most important words to consider to classify reviews, stemmed form. Source: [83].

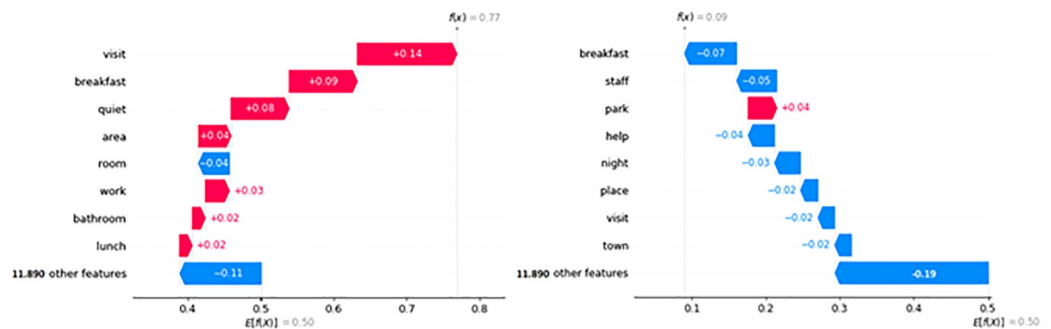


Figure 4.16: The Shapley values for two correctly classified reviews. On the left a positive-rated review and on the right a negative-rated review. Source: [83].

mostly influence the likelihood of a review to be positive. The absolute mean Shapley value, which is a measure of the words' impact on the model, shows that the vast majority of the available terms has a low, if no, impact on the model. In particular, using the elbow-point method [188] on the words' absolute mean Shapley values, 64 important features are determined, as shown in table 4.15

Figure 4.16 reports the words' Shapley values for two examples of reviews that are correctly classified by Random Forest, one with positive rating and the other with negative one. The positive review is mainly explained by the words *visit*, *breakfast* and *quiet*, that positively influence the review, representing positive aspects. On the contrary, the negative review shows a general dissatisfaction mostly affected by specific factors like *breakfast*, and *help*.

Chapter 5

Insights and future perspectives for startups and reviews' analysis in Industry 4.0

In this chapter the main conclusions and insights will be drawn about the application of graph theory and Machine Learning to the Complex Systems studied in the previous chapters. In particular, it will be first shown how graph modelling of the startup ecosystem, together with Machine Learning algorithms, can improve the understanding of its fundamental Complex System for Industry 4.0.

Then, the insights from the analysis of tourists' reviews about Apulian tourism offer will be discussed. In fact, since understanding consumer tastes and needs is of fundamental importance for a firm to be successful in the context of the Fourth Industrial Revolution and tourism is one of the most profitable businesses, this activity should be considered a pivot point to intercept and forecast tourists' demands.

5.1 Highlighting the best practices in innovation ecosystems through community detection

Section 4.1 describes an approach based on graph theory and community detection in order to obtain an equity-oriented rethinking of StartupBlink ranking about world countries' quality of their innovation ecosystems. As described in section 3.1.1, startups represent the main technological boost of the Fourth Industrial Revolution and their economic impact is still growing. Accordingly, identifying those factors that make a country successful in creating an enabling environment for startups will have a beneficial impact on its economy in terms of job creation, attraction of international investments funds and creation of technology monopolies.

Nonetheless, even if widely used, rankings do not take into account the dif-

ferent starting condition of the ranked elements and consequently, these tools should be used with care in making both political and business decisions. Accordingly, in this work, every country has been described by the corresponding set of World Development Indicators (WDIs) and a graph has been built, where countries are the nodes and a link is established between two countries according to the Pearson correlation of the corresponding WDIs. Then, community detection algorithms have been used to find sets of countries with the most similar development level, as described by the multi-faceted EWDI. If this is not a dataset representing relations, it could be regarded as such thanks to the correlation coefficient. This approach might be questionable, since simpler and more direct classical clustering algorithms could have been used to find sets of countries with the most similar WDIs: Means, K-Medoids and Hierarchical clustering [189]. Nonetheless, as shown in Appendix B.2, these methods are unsuitable and community detection is the most appropriate approach. In any case, graphs prove to be an invaluable help in retrieving hidden information in an unsupervised and purely data-driven way.

In particular, since WDIs cover multiple aspects of the social and economic performance of countries, it is not surprising that the network communities are characterized by different wealth classes but by an homogeneous wealth level therein. This result allows to relate the wealth level of a country with the quality of its innovation ecosystem and, above all, establish which countries need specific support or can be considered as examples of best practices in the technological innovation policies. The reliability of these results also rests on the robustness of the community detection. Both Spin Glass and Leiden algorithms give the same results.

The use of network communities as a tool to evaluate country performances is strengthened by a quantitative control to confirm the existence of a relation between community membership and expected ranking through the resolution ratio R . This parameter quantifies the tendency of the ranking-index distributions related to different network communities to be separated in a relevant way. The resolution ratio associated to the global StartupBlink index shows a good separation between communities, which allows to compare a country's performance with the expectation based on its wealth conditions.

Among the constituent indexes of StartupBlink, only one associated to a value $R > 1$ is the Business index, which measures the ease of business in the considered territory. This result is related to the presence among the WDIs, of indicators expressing the quality of bureaucratic practices and other aspects affecting the efficiency of firms. Instead, the Quantity and the Quality index seem not to be affected in a significant way by a territorial effect.

Deepening the analysis of the results, it can be seen that, unsurprisingly, the United States are a benchmark country both from a global and an ease-of-business point of view. In fact, conditions offered by the United States startup ecosystem to both entrepreneurs and investors are excellent. The United States ecosystem is focused in the New York and San Francisco areas. In particular, the technological center of the Silicon Valley represents the best choice to create products and initiatives that are viewed as appealing from the global market [190]. Instead, much of the United Kingdom's strength in the global startup ecosystem derives from its hub, London. In fact, in recent years, London has become the most successful startup ecosystem in Europe, with an ever growing number of startups, since it represents the first choice for fast growing US

startups willing to set up their European headquarters [191].

Furthermore it is worth noticing the role of Israel: it is, at the same time, a benchmark country for the global StartupBlink index and a *room-for-improvement* country for the Business index. This two-fold outcome can be useful for firms and investors. In fact, this indicates that, although Israel plays a leading role in the world of innovation ecosystems, its practices in boosting the startup environment should be improved. The apparent contradiction is related to the strong hierarchical nature of the Israel startup system: it has just a single dynamic innovation hub in Tel Aviv, while the rest of the territory does not reach comparable performances [192]. An independent confirmation of this last result is given by the fact that, since 2019, Israel has improved the quality of its socio-economic actions in order to boost the number of high-impact startups, as reported in [193]. This proves the ability of graph theory to unveil hidden patterns in data.

The approach developed in this work, based on a purely data-driven procedure, can represent the starting point to develop new objective methods for highlighting problematic scenarios and establish suitable policies in the innovation ecosystem.

As a further improvement, the proposed methodology can be employed to study the innovation ecosystems at a city level. Accordingly, for each country, it would be possible to identify the most successful local policies and characteristics in attracting startups and investments. This activity has the possibility to boost local economies.

5.2 Graph metrics and startups' success

In the previous section it has been discussed how graph modelling can be an invaluable help in correctly assessing countries' performances in enabling innovation ecosystems. Then, delving into the identification of the features characterizing the most successful startups (i.e. those collecting the greatest part of funds), datasets describing startups' funding rounds should be used. In this respect, Crunchbase is the most widely studied public dataset.

Correctly identifying those factors that make a young firm successful is of paramount importance both for investors and entrepreneurs. However, funds alone cannot give a complete view of what is happening in the startup ecosystem. In fact, they cannot answer the following questions: how many investors are involved in a funding round? How many funding rounds an investor is involved in? Are there elements with a high fund conveyance capacity? Even though answering these questions could not highlight the most successful elements, nonetheless they allow to determine the most strategic elements that let the entire ecosystem work. In order to extract these useful insights, a graph model for the dataset is required. In fact, graphs directly model the relationships among its elements and provide a set of metrics that quantitatively determine the role of each node in the system. In particular, since different metrics highlight different kinds of importance in the network, there are different types of strategic elements. Nonetheless, they are all defined as the *outliers* of the distribution of the corresponding metric distribution. Moreover, as described in section 4.2.1, some metrics can have a direct economic interpretation, so that their analysis can be easily interpreted by people unfamiliar with graph theory. This work

has developed a quantitative and easy-to-interpret model to account for the strategic importance of firms within the startup ecosystem. It has been also demonstrated (see section 4.2.2) that the information carried by graph metrics cannot be recovered using classical statistical analysis. Accordingly, this proves that graph representation of data is a fundamental tool for unveiling hidden information.

Then, it has also been demonstrated that a logistic regression model can be set up with which reliably forecasting the success of a firm up to five years in advance using only network metrics (see section 5.2.2). Specifically, the indegree has been identified as the most important centrality metric to predict whether a firm is an outlier of the distribution of collected funds, i.e. successful.

This study paves the way for future investigations, for example about the existence of a relationship between the investor types and the economic categories or between their country and that of the investors. In fact, the determinants of success for firms of different nationality or category are likely to be different.

5.3 Unveiling tourists' tastes and needs

In the previous two sections a fundamental Complex System for Industry 4.0 has been studied through graph theory in the startup ecosystem. This study has allowed to disclose its hidden patterns and highlight how the role of an element in this Complex System can be a proxy of its future success. Nonetheless, another feature of the Fourth Industrial Revolution is the possibility of consumers' interact through the sharing of their experiences using online social platforms. As underlined in section 1, this may influence other consumers' decisions. Accordingly, one key aspect in a firm activity is the understanding of consumers' needs and tastes. Since one of the most profitable economic activities, especially in Italy, is tourism (see section 3.2), it becomes critically important to highlight strengths and weaknesses of the tourism offer. This problem has been tackled in section 4.3 using TripAdvisor reviews about Apulian accommodation facilities.

In particular, a classification framework is shown that evaluates the rating and verbalization of the tourist experience and highlight its determinants to predict future satisfaction from the reviews.

Basically, it has been evaluated to which extent online reviews allow a reliable assessment of the tourists' experience and their satisfaction. First, it has been observed the presence of misleading reviews, i.e. case the numerical assessment did not match the sentiment expressed. Considering how the contradictory reviews are distributed among positive and negative reviews, been observed that 80% of negative reviews are contradictory. This preliminary analysis is essential for using reviews textual data to effectively forecast their rating. In fact, as described in section 4.3.2, Machine Learning algorithms are fed with balanced dataset of negative reviews and an equal number of randomly undersampled positive reviews. Accordingly, without the deletion of these contradictory reviews, these algorithms would have been fed with datasets having at least 40% of error level. Previous studies on the sensitivity of Machine Learning algorithms to noisy data show that model's accuracy decays almost linearly with the noise level. 40% of error level in data reduces by 30%-40% a model's accuracy [186, 194].

Consistently with previous studies, the proposed framework is able to cross-validate different models and evaluate their performance. Contrary to previous research [159], a cross-validated framework has been used, in order to get more robust results. In fact, since the results are independent from the train-test subdivision, biased results leading to inaccurate conclusions are avoided. Moreover, the results are comparable with those obtained using state-of-the-art deep learning methods [160].

In section 4.3.2 it is shown how the classification performance is robust independently on the adopted model or the specific considered performance metrics, even though Random Forest has the highest performance compared to the others. Moreover, it has been found a strong agreement with the predictions of the other models, especially SVM and XGB. This implies that the explainability analysis is independent on the particular considered model. This last analysis has been carried out using the Shapley paradigm, through which the models' decisions can be explained. In particular, this approach has been used for studying the decisions taken by Random Forest, the best performing model. On one hand, the findings underlined *what* are the the most important *words* related to places, meals and staff and in particular the word *breakfast*. This characterizes the typical tourist offer and can be explained by both the most common type of hospitality structures, namely bed-and-breakfast, and the connection with food, one of the most important elements of a tourist experience. On the other hand, the Shapley values also highlighted *how* these words affect the classification score. This helps in characterizing the experience and predicting the satisfaction (positive or negative evaluation).

These results have strong managerial implications in the way the tourist offer can be improved through the creation of personalized services on the basis of the reviews. Understanding the actual tastes and needs of reviewers through such behavioral-tracking data can unveil really valuable insights for business improvement and marketing effectiveness. Since consumers' tastes are dynamic and expensive to monitor, advances in the analysis tools can help in providing more useful information to enhance the offerings' quality and targets.

For the sake of simplicity, variables like nationality or age have not been taken into account. However, it is reasonable to assume that these factors can affect the judgements. Expectations and needs of a teenager are necessarily different from those of a family with children. Future studies will be devoted to enlarge the examined geographical area and take into account factors like age or nationality. Also, in this paper an ex post feature importance analysis based on Shapley values has been proposed, but it would be possible to consider an ex-ante feature importance step in the learning phase. The design and implementation of dedicated strategies to maximize and exploit the informative content provided by online reviews deserves further investigations.

Although the main aim of this work is to analyze tourists' reviews and give useful insights to tourism stakeholders, the proposed framework could be applied as a general framework in all the analyses involving textual data (e.g. Amazon products' reviews, events logs). By analyzing products and events' reviews this model helps highlighting those aspects that mostly influence reviewers' feelings. In fact, the main components of the proposed workflow are: (1) *Review scraping*, obtained by using packages that are freely available to every programming language [195]. These packages can be used to scrape almost all social media platforms (like Booking.com, Facebook, Amazon) and ob-

tain the desired text (2) *NLP techniques* and *Sentiment Analysis* that have achieved optimal standard in analysing textual data and in extracting useful insights about the meaning of texts [196]. *Machine Learning* and *Explainability* algorithms that are widely used in fields like wildfire preventions [197, 198], medicine [105, 199] and drug discovery [200].

5.4 Future perspectives of Complex Systems and Machine Learning in Industry 4.0

This work deals with two main arguments: (1) the modelling of the startup ecosystem, a Complex System whose importance for Industry 4.0 is often overlooked, in order to find the success keys for both firms and investors through Machine Learning algorithms; (2) the use of NLP techniques and Machine Learning for the analysis of unstructured textual data in order to extract insights about consumers' tastes and needs. Certainly, Complex Systems and Machine Learning applications are not limited to these two cases and many are the leading examples on which the combined use of Machine Learning and graph theory can be beneficial for extracting insights not otherwise available and forecasting systems' evolution.

For example, the analysis of unstructured textual data, which has been carried out through the TF-IDF matrix, can be accomplished using Deep Learning tools particularly suited for dealing with sequences of words like text and time series [201, 202]. In fact, the so-called *Recurrent Neural Networks* implement a *memory* mechanism through which an input is treated taking into account the previous ones. Since text is a stream of subsequent words and their ordering is important for its understanding, these tools will enhance the textual analysis with respect to the TF-IDF model, in which the ordering of words is discarded (this last approach is known as *bag-of-words* method). Moreover, as underlined in section 1.1, Industry 4.0 is characterized by an unprecedented amount of data coming from all productive elements (equipped with sensors) of a firm. These data are, for the greater part, in the form of time series [203]. As a consequence, improving the analysis of time series, for both regression and classification purposes, will benefit the analysis of firms' activity (e.g. supply chain optimization) and help in unveiling the corresponding most important features. Accordingly, the next research steps will delve into the use of these novel Recurrent Neural Networks for the analysis of time series, in general, and of textual data in particular.

Moreover, the methods shown in this work can be combined in a single framework for improving *Recommender Systems*. Recommender Systems are algorithms that, depending on the particular application, suggest a user what to follow, watch or buy based on the user's history and biographical information [204]. Many online services (e.g. Amazon, Netflix) use these instruments to enhance the user experience. Namely, these tools are up to now, based on users' personal information, user-user similarities and suggested items' likeness [205]. Nonetheless, an improvement on their performance can come from considering relations among users. In fact, users can be linked by friendship or follow relations in an online social network like Facebook or Tencent Weibo. Since it seems wise to assume that these relations link people with similar in-

terests and tastes, taking them into account will improve the performance of recommender systems that are often penalized by having a great number of items to suggest [206]. Directly taking into account relationships among users into Recommender Systems was not possible until the introduction of particular *Artificial Neural Network* models that combine the users' features (e.g. tastes) with a graph modelling of their relations (the *Graph Neural Networks* (GNNs) [207]). These neural networks are a recent and very active research branch [208]. GNNs have proven to significantly improve classical models in different fields like traffic forecast, book Recommender Systems, web page classification [209]. Accordingly, the next steps in the application of graph theory and Machine Learning to Industry 4.0 will deal with this new promising tool.

Conclusion

The Fourth Industrial Revolution is radically changing firms' business models, production methods and interactions with consumers, driving their transition to *smart firms* [210]. Accordingly, this phenomenon is still an active research area for different knowledge fields like electronics, computer science and economics [211,212,213]. One of the aims of this work is to shed light on the often overlooked relations of Industry 4.0 with Complex Systems [214]. In particular, this study focuses on its fundamental link with startup ecosystems, that are Complex Systems comprising startups, their funders and the corresponding funding relations [105]. In fact, startups are usually the main boosts of technological innovation, which represents firms' main obstacle to the transition towards the Industry 4.0 paradigm [142]. The fundamental role of Industry 4.0-related startups is evidenced by their economic value: reach 200 billion euros in the next two years and will triple by 2030 [33]. Accordingly, establishing an effective innovation ecosystem will boost Industry 4.0 and have a positive impact on countries' economy. From these considerations follow the first research question stated in the Introduction:

- **RQ-1** Given the importance of startups in boosting countries' economies, how is the effectiveness of a country's innovation ecosystem influenced by the socio-economic context in which it grows?

To answer this question, it has been first acknowledged that the quality of a country's startup ecosystem is usually expressed through rankings comparing countries' achievements in supporting innovation [104]. Nonetheless, rankings offer a too limited view of the status-quo since they do not consider the socio-economic conditions underpinning a country's position in the ranking. Moreover, since they are usually built using arbitrary weighted averages of indices, they are prone to be contested. Despite these problems, rankings are nowadays widely used to take political and business decisions [123]. Then, it becomes of paramount importance to determine an equity-oriented rethinking of rankings. The analysis carried out in this work is the first quantitative and data-driven attempt of this kind in the startup ecosystem [122]. It will be beneficial for both public and private stakeholders. In particular, one of the most considered public rankings about countries' startup ecosystem has been analyzed: StartupBlink [192]. It has been highlighted how a graph model taking into account the multi-faceted socio-economic background of countries, expressed by the World Development Indicators given by the World Bank, insights on this ranking. Comparing the results of community detection with countries' positions in StartupBlink, it has been possible to quantitatively highlight both the problematic scenarios hidden in highly-ranked countries and the

unexpressed potentials of low-ranked countries. For example, it has been shown that, even though India and Brazil belong to the community of low-income countries, they overperform their community peers in setting up an effective innovation ecosystem, reaching the levels of high-income countries. These countries' great innovation potential would have been completely overlooked if only the ranking had been considered. Moreover, the case of Israel deserves attention with- standing its high position in the ranking, an objective criticality emerges in its easiness of doing business. These results have been found in a completely data-driven way and are confirmed a posteriori by countries' government mea- sures [193]. It should be underlined that graph modelling has not been simply an option, but the only way to obtain such insights. In fact, as shown in this work, classical clustering methods have proved to be unfit for linking the socio- economic context of countries with their outcome in the ranking.

Then, this study goes into a greater level of detail about the functioning of a startup ecosystem. In fact, given its importance for Industry 4.0, it seems wise to pinpoint which elements are the most strategic for this ecosystem, what are their characteristics and which features make a startup successful. The second research question follows:

- **RQ-2.** Who are the most strategic elements in a startup ecosystem? Is there a relation between the strategic value of a startup in this system and its future success?

These findings would be beneficial for entrepreneurs, funders and the devel- opment of Industry 4.0. In order to answer these questions, this work focused on one of the most cited and studied databases about startup funding: *Crunchbase*. The importance of a startup is usually measured in terms of the funds it is able to collect [105] and, moreover, this is the only quantitative information con- tained in *Crunchbase*. Nonetheless, this does not provide a complete overview on the functioning of the startup ecosystem. In fact, just considering funds does not allow, for example, to highlight the role of those elements characterized by a high money conveyance, they are crucial for the working of the ecosystem but are completely overlooked by a naive funds analysis. On the contrary, as shown in this work, graph theory gives the opportunity to use ad-hoc metrics to quantitatively define different kinds of importance, or strategic value, of the elements. In particular, the features characterizing different kinds of investors and startups have been found. For example, the role of *Accelerators* has been highlighted. They are private funders that, above all, mentor startups in all their activities and following this work's analysis, they are characterized by two properties: (1) they are able to attract a higher number of funders than all the other investors; (2) they are the most effective in conveying money within the startup ecosystem. Moreover, it has been possible to highlight also the fea- tures characterizing different kinds of startups. For example, those dealing with *e-Payments* are not characterized by raising high funds, but by their ability of investing and conveyance. These insights are confirmed by the economic litera- ture [180]. These insights have been found in a purely data driven way thanks to graph modelling and could not have been found using naive statistical analysis of funds.

Moreover, the role of a startup in the ecosystem, as measured by graph metrics, has been linked to its future success, defined as the ability of collecting

more funds than all the others. Even though this definition of success is not the only possible one, it is both straightforward and the most common in the economic literature [215]. It should be underlined that this last analysis is independent from firm business characteristics (e.g. number of employees, headquarter's country) that are also difficult to obtain but is based only on the funding relations defining the startup ecosystem. It has been found that the role a startup plays in the startup ecosystem is a good proxy of its ability of collecting high funds within four years.

Another feature characterizing Industry 4.0 is consumers' possibility of sharing their opinions and ideas about products and experiences by posting reviews on online social platforms (e.g. TripAdvisor, Facebook, Amazon). These reviews can in principle influence other consumers spread all over the world in their decision making process and determine the success of this phenomenon is so well acknowledged and fundamental in the Industry 4.0 context such that consumers are also identified as *co-producers* [17]. Accordingly, one of the key aspects of a firm's activity should be that of intercepting and forecasting consumers' needs and tastes from reviews in order to have more targeted marketing campaigns and a more effective production. From these considerations stems the third research question of this work:

- **RQ-3.** How can insights from textual data be automatically extracted?

Accordingly, this work focused on extracting insights from reviews about one of the most profitable economic activities, especially in tourism [216]. In particular, reviews extracted from TripAdvisor about Apulian tourism accommodation facilities have been considered. Thanks to Natural Language Processing (NLP) techniques and Machine Learning tools, it has been possible to highlight those concepts that mainly influence reviews in being positive or negative. Specifically, thanks to Explainable Machine Learning (XAI) techniques based on Shapley values [92], words related to *food* and *staff* can be highlighted as positively perceived from tourists, meaning that food quality and a friendly staff are fundamental to be appreciated by tourists. On the other hand, words like *room* and *work* make reviews being negative. This means that room quality and services needed for smart working are aspects that should be improved for an effective tourist welcome. These insights have been found feeding reviews' textual data into Machine Learning algorithms combined with Explainability techniques, without any other external additional information.

The analyses carried out in this work show that shedding light on the relations among Industry 4.0, Complex Systems and Machine Learning gives the possibility of extract useful insights for boosting firms' competitiveness and innovative potential. Moreover, some future perspectives of this work can be pointed out. As regards RQ-1, it can be stated at a local level the socio-economic context of a region (resp. a city) influence the effectiveness of innovation ecosystems. Using a wide set of local socio-economic indicators, a graph model taking them into account can be built and insights can be derived by comparing graph's outcomes from community detection with those deriving from rankings of local innovation ecosystems like that of *Startup Genome* [119] or the corresponding local version of *StartupBlink* [178]. Accordingly, it will be possible to establish more targeted cues about which political/economical actions to take for exploiting the unexpressed potential of territories and boost their economy.

As regards the second research question, graph theory has proven to be effective in measuring different kinds of importance of an element in the startup ecosystem giving the possibility of deriving insights about the most strategic elements of this system. Moreover, a Logistic Regression model has been built that relates the metrics' values of an element with its probability of being successful in the future. This analysis can be deepened in two ways: considering more sophisticated Machine Learning models (e.g. Random Forest, Neural Network architectures); (2) adding firms' business information (e.g. number of employees, headquarter's nation) as input features of the model and determine if the model improves its performance or not and what are the features leading a firm towards success.

Considering RQ-3, this work has been focused on the analysis of tourists' reviews about Apulian accommodation facilities. Even though the proposed framework has proven to be effective in analyzing textual data, it can be improved by using the tourists' characteristics (e.g. nationality, age) as input features in the Machine Learning models aiming at determining reviews' sentiment. In fact, the needs and tastes of teenagers are likely to be different from those of older people and, accordingly, the evaluation of tourism offer would also differ. Moreover, the developed framework can be applied to every activity in which the automated analysis of textual data is fundamental for extracting insights about products or experiences. Nevertheless, there is still room for improvement. In fact, the NLP techniques used are based on the so-called *bag-of-words* model, according to which the words alone are able to express the meaning of a message, regardless of their order. Accordingly, the considered Machine Learning models use every single word as an input feature (through the TF-IDF matrix). Even though this analysis gives positive results, it can be improved in two respects: (1) using more recent Neural Network architectures like the Multi Layer Perceptron; (2) by leveraging the order in which words appear in a review and considering models that are able to use this additive information. Accordingly, the next steps of the analysis done for answering RQ-3 would encompass the use of Neural Network architectures endowed with *memory* mechanisms, like the Recurrent Neural Network (e.g. Long-Short Term Memory Gated Recurrent Unit), in order to be ever more precise in intercepting customers' needs and tastes and boosting firms' competitiveness.

Appendices

Appendix A

2019 StartupBlink ranking

In Table A.1 the 2019 StartupBlink country ranking is reported, with the overall score and the three indexes that compose it. The first two columns indicate the country names and the corresponding ISO-3166 alpha-3 codes. The remaining four columns represent the StartupBlink index and the corresponding constitutive indices (Quantity, Quality and Business index) respectively. Continues on the next page.

Country	ISO code	StartupBlink	Quantity	Quality	Business
United States	USA	44.09	12.29	22.02	9.78
United Kingdom	GBR	16.72	1.86	5.10	9.76
Canada	CAN	15.87	1.24	5.10	9.54
Israel	ISR	14.63	0.35	5.21	9.07
Australia	AUS	12.95	0.64	2.71	9.61
The Netherlands	NLD	12.91	0.34	3.27	9.29
Sweden	SWE	12.77	0.19	2.87	9.71
Switzerland	CHE	12.53	0.21	3.06	9.26
Germany	DEU	12.46	0.71	2.25	9.50
Spain	ESP	12.40	0.56	2.42	9.4
France	FRA	11.45	0.50	1.59	9.36
Finland	FIN	11.37	0.11	1.63	9.62
Estonia	EST	11.27	0.10	1.52	9.64
Ireland	IRL	11.12	0.16	1.44	9.52
Russia	RUS	10.88	0.71	1.18	8.98
Denmark	DNK	10.66	0.14	0.65	9.87
India	IND	10.65	1.48	0.59	8.58
Lithuania	LTU	10.52	0.10	0.74	9.67
South Korea	KOR	10.47	0.07	0.97	9.43
Poland	POL	10.45	0.21	0.89	9.35
Singapore	SGP	10.43	0.06	0.89	9.48
Czech Republic	CZE	10.17	0.10	0.75	9.31
Japan	JPN	10.10	0.14	0.72	9.24

Country	ISO code	StartupBlink	Quantity	Quality	Business
Belgium	BEL	10.09	0.13	0.81	9.14
Italy	ITA	10.07	0.31	0.72	9.03
New Zealand	NZL	10.06	0.06	0.10	9.90
China	CHN	10.04	0.42	1.30	8.32
Austria	AUT	10.04	0.12	0.46	9.47
Portugal	PRT	10.03	0.17	0.53	9.33
Chile	CHL	9.77	0.18	0.64	8.95
Ukraine	UKR	9.72	0.23	0.81	8.69
Mexico	MEX	9.68	0.18	0.52	8.98
Thailand	THA	9.67	0.11	0.51	9.05
Colombia	COL	9.45	0.15	0.51	8.79
Bulgaria	BGR	9.26	0.08	0.28	8.89
Serbia	SRB	9.21	0.07	0.06	9.09
Brazil	BRA	9.21	0.46	0.72	8.03
Romania	ROU	9.21	0.13	0.07	9.02
Hungary	HUN	9.18	0.08	0.10	9.00
United Arab Emirates	ARE	9.12	0.11	0.08	8.93
Indonesia	IDN	8.89	0.10	0.54	8.25
Greece	GRC	8.82	0.09	0.06	8.67
Turkey	TUR	8.65	0.22	0.05	8.37
Argentina	ARG	8.63	0.16	0.61	7.85
Latvia	LVA	8.48	0.05	0.34	8.09
Norway	NOR	8.41	0.05	0.07	8.30
Malaysia	MYS	8.38	0.10	0.51	7.76
Slovenia	SVN	7.91	0.04	0.15	7.72
Slovakia	SVK	7.80	0.05	0.06	7.69
Croatia	HRV	7.57	0.06	0.10	7.41
South Africa	ZAF	7.55	0.05	0.51	7.00
Kenya	KEN	7.42	0.06	0.01	7.36
Luxembourg	LUX	6.99	0.03	0.38	6.57
Philippines	PHL	6.82	0.04	0.50	6.27
Belarus	BLR	6.33	0.03	0.02	6.28
Nigeria	NGR	6.00	0.11	0.00	5.89
Peru	PER	5.80	0.05	0.51	5.24
Iceland	ISL	5.66	0.02	0.29	5.35
North Macedonia	MKD	5.64	0.02	0.07	5.54
Egypt	EGY	5.60	0.06	0.00	5.54
Pakistan	PAK	5.34	0.08	0.00	5.26
Georgia	GEO	5.16	0.01	0.04	5.11
Armenia	ARM	5.09	0.02	0.06	5.01
Rwanda	RWA	4.74	0.01	0.01	4.71
Morocco	MAR	4.70	0.02	0.00	4.68
Moldova	MDA	4.45	0.01	0.04	4.40
Azerbaijia	AZE	4.41	0.01	0.02	4.38
Cyprus	CYP	4.39	0.01	0.15	4.23
Kazakhstan	KAZ	4.35	0.01	0.01	4.33
Puerto Rico	PRI	4.16	0.01	0.04	4.11
Uruguay	URY	4.15	0.02	0.05	4.07

Country	ISO code	StartupBlink	Quantity	Quality	Business
Vietnam	VNM	4.06	0.02	0.32	3.72
Jordan	JOR	3.96	0.02	0.02	3.91
Tunisia	TUN	3.86	0.01	0.01	3.83
Ghana	GHA	3.77	0.02	0.01	3.74
Bosnia and Herzegovina	BIH	3.72	0.01	0.04	3.67
Ecuador	ECU	3.62	0.02	0.01	3.59
Sri Lanka	LKA	3.61	0.02	0.01	3.58
Dominican Republic	DOM	3.47	0.01	0.01	3.45
Saudi Arabia	SAU	3.35	0.02	0.00	3.32
Uganda	UGA	3.03	0.01	0.00	3.02
Lebanon	LBN	2.80	0.01	0.03	2.76
Iran	IRN	2.72	0.02	0.00	2.70
Cameroon	CMR	2.61	0.01	0.01	2.60
Albania	ALB	2.38	0.00	0.05	2.33
Costa Rica	CRI	2.29	0.01	0.03	2.26
Bangladesh	BGD	2.22	0.02	0.00	2.20
Jamaica	JAM	2.17	0.01	0.05	2.12
Malta	MLT	2.07	0.01	0.10	1.96
Botswana	BWA	1.98	0.00	0.05	1.93
El Salvador	SLV	1.97	0.00	0.02	1.94
Zambia	ZMB	1.92	0.00	0.01	1.91
Venezuela	VEN	1.82	0.01	0.00	1.80
Trinidad and Tobago	TTO	1.62	0.00	0.03	1.60
Bahrain	BHR	1.61	0.00	0.07	1.54
Paraguay	PRY	1.49	0.01	0.02	1.46
Qatar	QAT	1.24	0.01	0.06	1.18
Bolivia	BOL	0.82	0.01	0.01	0.80
Algeria	DZA	0.81	0.00	0.00	0.80
Ethiopia	ETH	0.01	0.01	0.00	0.00

Appendix B

StartupBlink community detection and clustering analyses

B.1 Spin Glass and Leiden algorithm feature space exploration

The heatmaps in Figures B.1–B.10 represent the performance indicators of the Spin Glass and Leiden community detection algorithms at these different steps required to identify the most stable and reliable partition.

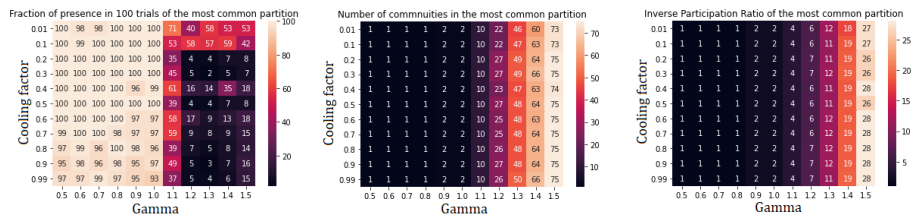


Figure B.1: Performance indicators for Spin Glass algorithm applied to the whole StartupBlink network. Two communities are found: community 0 (49 nodes) and community III (51 nodes).

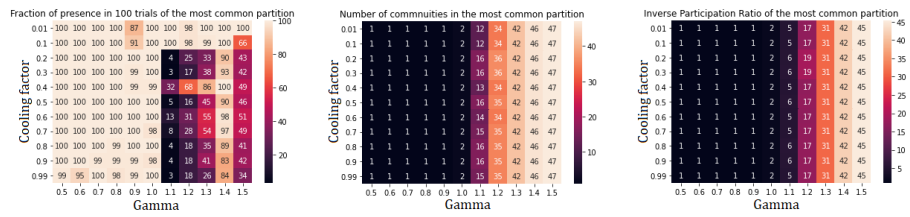


Figure B.2: Performance indicators for Spin Glass algorithm applied to community 0. Two subcommunities are found: community I (22 nodes) and community II (27 nodes).

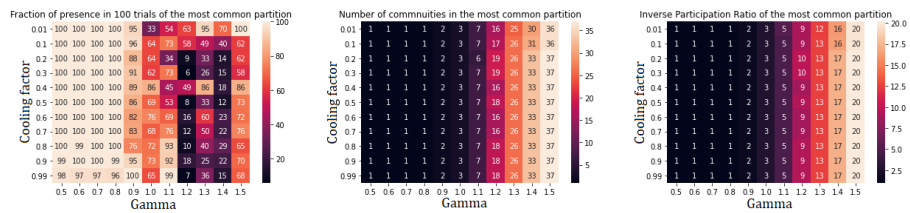


Figure B.3: Performance indicators for Spin Glass algorithm applied to community III. There is no further subdivision.

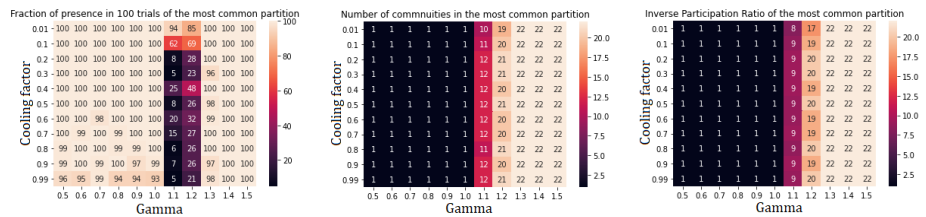


Figure B.4: Performance indicators for Spin Glass algorithm applied to community I. There is no further subdivision.

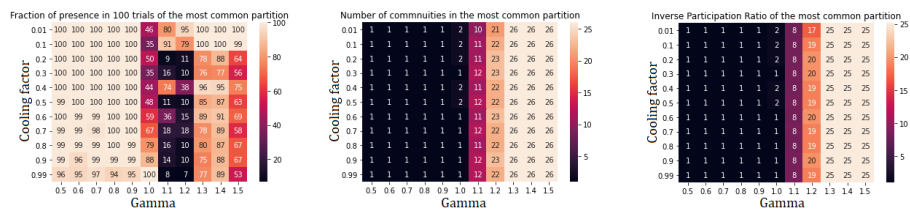


Figure B.5: Performance indicators for Spin Glass algorithm applied to community II. There is no further subdivision.

B.2 Why not classical clustering methods?

The methods exposed previously to obtain groups of similar countries in an unsupervised way are based on graph theory. Nonetheless, it may be questionable the use of graph methods to model the StartupBlink countries, represented by numerical features, the WDIs, that do not describe interactions

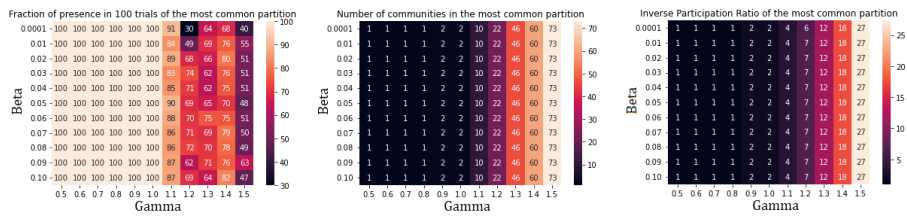


Figure B.6: Performance indicators for Leiden algorithm applied to the whole StartupBlink network. Two communities are found: community 0 (49 nodes) and community III (51 nodes).

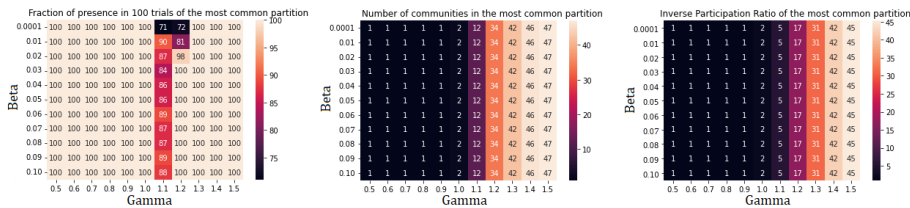


Figure B.7: Performance indicators for Leiden algorithm applied to community 0. Two subcommunities are found: community I (22 nodes) and community II (27 nodes).

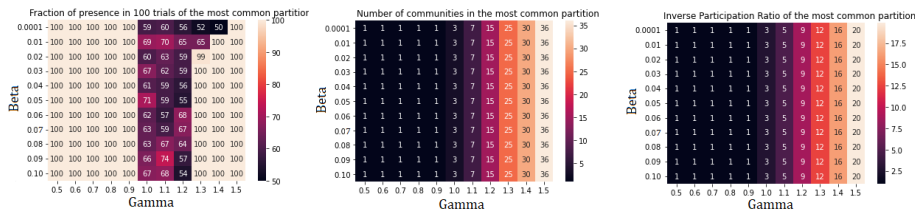


Figure B.8: Performance indicators for Leiden algorithm applied to community III. There is no further subdivision.

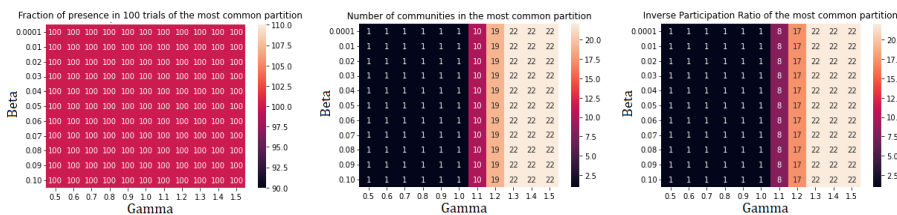


Figure B.9: Performance indicators for Leiden algorithm applied to community I. There is no further subdivision.

among them, but simply their characterising features.

Actually, the so called *clustering methods* have been developed to find groups (also called *clusters*) of similar objects in a set, representing data associated to each object as points in a multidimensional space, the *feature space* [44].

In order to show the advantage of the network model, even with this kind

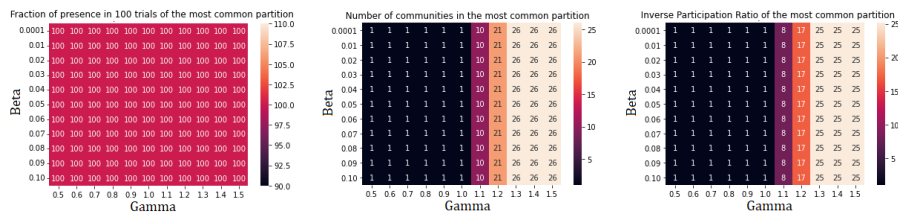


Figure B.10 Performance indicators for Leiden algorithm applied to community II. There is no further subdivision.

of data, different classical clustering algorithms have been applied, comparing their performances to those obtained with community detection algorithms. In particular, three of the most relevant algorithms in clustering problems have been considered: *k-means*, *K-medoids* and *hierarchical clustering* [217, 46].

K-means is one of the most popular clustering algorithms widely used in both academic and industrial settings [144]. This algorithm focuses on the minimization, through an iterative process, of the *sum of squared errors* (SSE), determined by using the Euclidean distance among points:

$$SSE = \sum_{j=1}^K \sum_{i \in S_j} \|x_i - \mu_j\|^2$$

where $j = 1, \dots, K$ identifies the objects, K is the number of clusters, S_j the j -th cluster, x_i is the data vector corresponding to the i -th object, μ_j the centroid of the j -th cluster and $\|\dots\|$ denotes the Euclidean norm.

K-means has two drawbacks: (1) it is a stochastic algorithm, in which different runs generally provide different clustering results; (2) the number of clusters, K , should be fixed a priori.

To solve the first issue, 100 different runs of *K-means* have been performed, in order to check the robustness of the minimization process. As regards the second issue, the optimal number of clusters can be found considering both the SSE and the mean *Silhouette score* [218] together. The latter is a measure of the clustering quality, based on averaging over all objects the *Silhouette score*, defined for a given data vector a_i as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

where a_i is the average distance between a_i and all other points in the same cluster, while b_i is the average distance between a_i and all points in the nearest cluster. The optimal number of clusters corresponds with the *elbow point* of the SSE vs K curve [219] and, at the same time, with the maximum of the mean *Silhouette score*. If these two conditions are not satisfied together, it can be concluded that *K-means* is not well suited for clustering the considered data.

The same reasoning on the clustering quality applies to *K-medoids*. This algorithm is similar to *K-means* in which actual data points are chosen as cluster centers, rather than the centroids. Moreover, *K-medoids* can be used with arbitrary distances [220] in order to calculate SSE and the mean *Silhouette score*. In this work, three common metrics have been used: Euclidean, Cosine and Manhattan.

Another approach is the *Hierarchical clustering* applied to the data points in the feature space. This class of algorithms differs from previous ones since it does not require to fix the number of clusters. In this work, a Hierarchical clustering algorithm is implemented that starts with each data point considered as an individual cluster, and iteratively merges the closest pairs of clusters until it ends up with a single cluster encompassing all data points.

In order to avoid the effect of outliers and putting all data points in clusters on a same ground, the *average linkage* is applied (see section 4.2.1). In this case, the Euclidean, Cosine and Manhattan metrics are considered to calculate distances among clusters.

Moreover, Hierarchical clustering, unlike K-Means and K-Medoids, is deterministic and produces *dendrograms* which can be helpful in interpreting the results. It is important to remark that, since Hierarchical clustering algorithms are not optimization problems, SSE and Silhouette are not reliable measures of the partition quality. Accordingly, the IPR values at various levels of the dendrogram can be used as a factor to evaluate the quality of the subdivision for each of the considered metrics.

B.2.1 Clustering results for StartupBlink countries

In this section the performance of classical clustering algorithms will be shown. In particular, it will be observed that the performance of classical clustering algorithms is not satisfactory, thus making network methods necessary.

In figure B.11, the SSE and mean Silhouette score of the K-means algorithm are presented, as a function of the number of clusters (K). It can be observed the absence of an elbow-point in the SSE plot. Moreover, the maximum mean Silhouette value is obtained for $K = 2$, where SSE also reaches its maximum. This implies that K-means is not well suited for an efficient partition of StartupBlink countries.

In Figure B.12, one can observe the same inconsistency in the case of SSE and mean Silhouette for K-medoids, with the Euclidean, Cosine and Manhattan metrics. Therefore even K-medoids algorithms cannot be considered as a suitable clustering method for StartupBlink countries.

As regards hierarchical clustering algorithm, figure B.13 shows the corresponding dendrogram. Moreover, the IPR values of the various partitions returned by the algorithms are considered as a measure of the clustering quality.

In Table B.1, the IPR values are shown corresponding to a number of clusters going from $K = 10$ to $K = 2$. It can be noticed a discrepancy for all values of K , between the number of groups and the IPR, indicating the presence of clusters with a very small number of elements.

Actually, this tendency to create highly uneven partitions can be already observed by inspecting the dendrograms of figure B.13. On the other hand, such a fragmentation is avoided in the network community detection, as demonstrated both by the final (22, 27, 51) partition reported in section 4.1.3, and by the detailed results of the community detection algorithm (see figures in the previous section), where at each step, the optimal communities are characterized by IPR close to the partition cardinality.

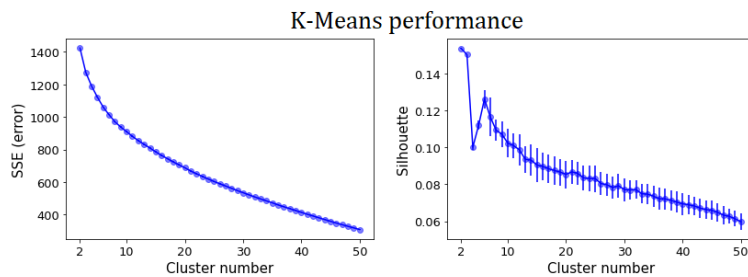


Figure B.11: SSE (left panel) and mean Silhouette value (right panel) of means clustering for StartupBlink countries, at different values of K (numbers of clusters). Error bars are determined by the variance of the considered quantities over 100 runs of the algorithm.

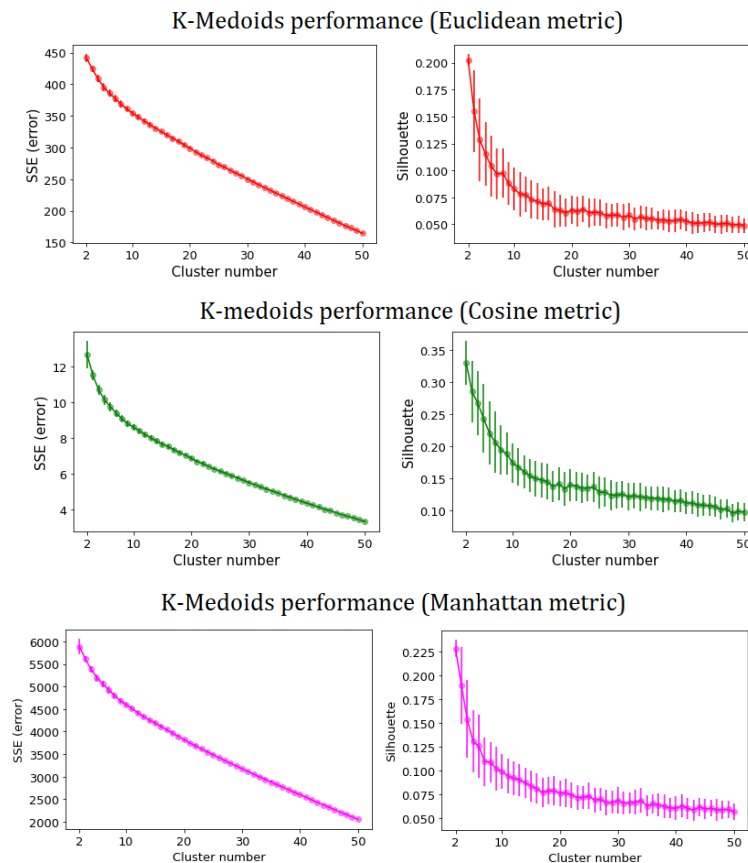


Figure B.12: SSE (panels in the left column) and mean Silhouette value (panels in the right column) of K-medoids clustering for StartupBlink countries, at different values of K , for Euclidean, Cosine and Manhattan metric. Error bars are determined by the variance of the considered quantities over 100 runs of the algorithm.

Table B.1: IPR values of the partitions returned by hierarchical clustering algorithms based on the Euclidean, Cosine and Manhattan metrics, different cluster numbers k .

k	Euclidean	Cosine	Manhattan
10	1.918	1.927	3.030
9	1.905	1.916	2.883
8	1.889	1.903	1.468
7	1.880	1.879	1.467
6	1.368	1.879	1.433
5	1.252	1.869	1.423
4	1.062	1.337	1.300
3	1.062	1.224	1.299
2	1.020	1.173	1.041

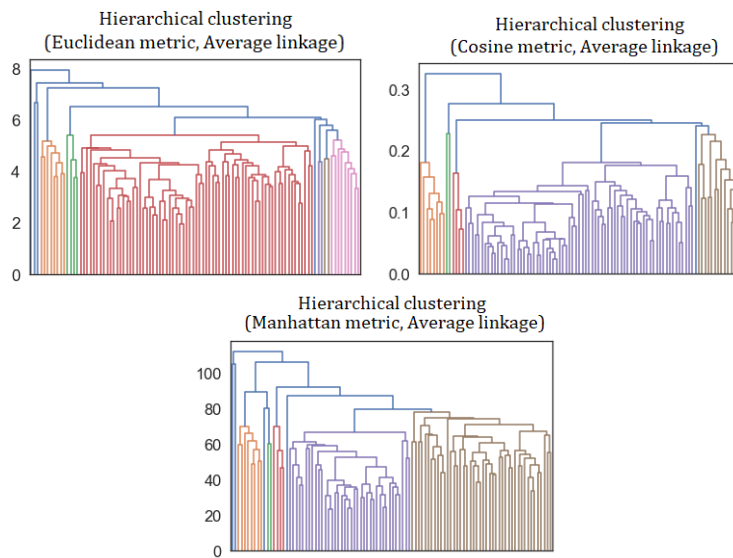


Figure B.13 Hierarchical clustering dendrograms, obtained using the Euclidean (left panel), Cosine (center) and Manhattan (right) metrics with the average linkage method. The vertical axes report the values of the metric.

Appendix C

Crunchbase main features and statistical analyses

C.1 Dataset description

In Table B.1 the 17 different datasets composing Crunchbase are listed, together with their brief description.

Table C.1

File Name	Short Description
1. Acquisitions	Data about acquisitions
2. Category_groups	A list of all economic categories within the data
3. degrees	Educational qualification of tracked people
4. event_appearances	events and participating people
5. events	A list of all recorded events
6. funding_rounds	Description of funding rounds
7. funds	The file includes all present investment funds
8. investment_partners	partnerships established in funding rounds
9. investments	Information about leader investors in funding rounds
10. investors	A description of all Crunchbase investors
11. ipos	Firms at initial public offering stage
12. jobs	Job career of tracked people
13. org_parents	The list of subsidiaries and controller companies
14. organization_descriptions	Description of firm activities
15. organizations	A detailed description of all Crunchbase firms
16. people	A list of all people in Crunchbase
17. people_description	A description of tracked people

C.2 Most present elements' attributes

Table B.2 shows the most present features of Crunchbase dataset.

Table C.2

Ranking	Nationality	Economic category	Investor type
1.	USA (53.6%)	Internet services (19.3%)	Business Angel (60.4%)
2.	UK (7.6%)	e-Payments (14.4%)	Venture Capital (27.8%)
3.	IND (4.2%)	Software (6.1%)	Private equity (6.2%)
4.	CAN (3.0%)	Science (5.8%)	Accelerator (1.9%)
5.	CHI (2.9%)	ICT (5.6%)	Government Office (1.1%)
6.	DEU (2.8%)	e-Commerce (5.0%)	Incubator (1%)
7.	FRA (2.3%)	Sharing transportation (4.4%)	Investment bank (0.9%)
8.	ISR (1.7%)	Apps development (4.3%)	Fund (0.5%)
9.	AUS (1.5%)	Healthcare (4.1%)	Secondary purchaser (0.03%)
10.	ESP (1.3%)	Advertising (3.9%)	Startup competition (0.003%)

C.3 Funding and network metrics global distribution differences and top fifty ranking

Normalized distributions of Funding, Indegree, Outdegree and Betweenness are shown in figure C.1. All network centrality distributions are significantly different from the funding on degree ($p \sim 10^{16}$), Outdegree ($p \sim 10^6$), Betweenness ($p \sim 10^6$).

Figure C.1

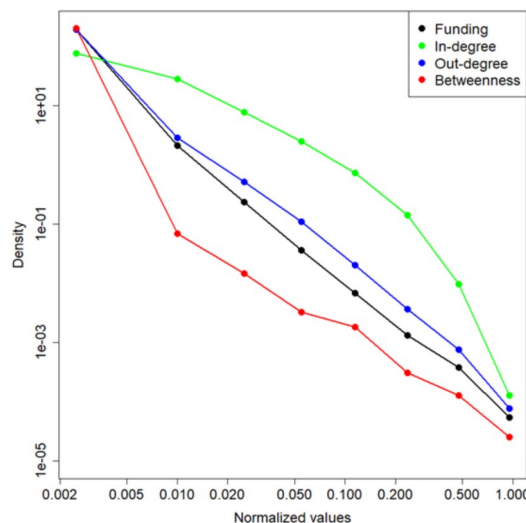


Table C.3

Rank	Funding	Indegree	Outdegree	Betweenness
1.	Verzion Comm.	Uber	500 Startups	Y Combinator
2.	Tsinghua Unigr. Int.	Atrium LTS	Y Combinator	FundersClub
3.	Didi Chuxing	Flexport	Sequoia Capital	Techstars
4.	Tesla	DocuSign	New Enterprise Associates	StartX
5.	China Unicom	Pinterest	Intel Capital	Alibaba
6.	Uber	SeatGeek	Accel Partners	Alchemist Accel.
7.	Rosneft	Opendoor	NYSERDA	Groupon
8.	WeWork	CardioDx	Kl. Perk. Cauf. & Byers	Salesforce
9.	AT&T Wir. Mob. Gr.	Lyft	SOSV	Google
10.	Alibaba	Prosper	Wayra	Crowdcube
11.	Meituan-Dianping	Fab	Draper Fisher Jurvetson (DFJ)	Seedcamp
12.	Flipkart	Mattermark	SV Angel	Betaworks
13.	Clearwire	Active Network	Start-Up Chile	Startupbootcamp
14.	Hilton Worldwide	TransMedics	Bessemer Venture Partners	Seedrs
15.	Apple	Tesla	Techstars	WR Hambrecht
16.	SH Pudong Dev. Bank	Practice Fusion	Right Side Cap. Manag.	Baidu
17.	Sberbank	Domo	Technology Development Fund	DST Global
18.	COFCO	Neuronetics	Greylock Partners	AngelList
19.	Jumpstart Ltd	PTC Therapeutics	First Round Capital	500 Startups
20.	Charter Comm.	Airbnb	Goldman Sachs	AOL
21.	Ping An	EndoGastric Sol.	Index Ventures	Tencent Hld.
22.	Suning	ecomom	Lightspeed Venture Partners	Digital Curr. Gr.
23.	Ant Financial	Artsy	Battery Ventures	Slack
24.	Airbnb	Bluesmart	Plug and Play	Amplify.LA
25.	Gas Natural	Scopely	High-Tech Gruenderfonds	Yahoo
26.	Nvidia	Namely	Crowdcube	Visionplus
27.	Evonik Industries	Pivot3	Brand Capital	Rock Health
28.	First Data Corp.	Sun Basket	Venrock	OurCrowd-GCai
29.	Grab	Keen IO	Andreessen Horowitz	Didi Chuxing
30.	Ele.me	Memebox Corp.	Benchmark	JFDI.Asia
31.	AccorHotels	Klout	General Catalyst	CircleUp
32.	Xerox	Boxed	Khosla Ven.	Amazon
33.	Allegro	Spotify	Norwest Ven. Ptrs - NVP	Anthemis Group
34.	Toys R	Meru Networks	GV	Cisco
35.	Toutiao	Doppler Labs	Redpoint	Kickstarter
36.	Ola	Casper	Menlo Ventures	Uber
37.	Reliance Jio Inf. Ltd.	Kamcord	Canaan Partners	Xiaomi
38.	B2M Solutions	Proterra	Atlas Venture	Lighter Capital
39.	Magic Leap	Actelis Networks	Northstar Ventures	SeedInvest
40.	Roche	Luxe	Matrix Partners	Entrepreneur First
41.	Lazada Group	Calient Tech.	Pol. Partners	LetsVent.
42.	Snap Inc.	Slack	U.S. Venture Ptrs (USVP)	Garage Tech. Ven.
43.	Lyft	GENBAND	Seedrs	PayPal
44.	Safaricom	Black Duck Sw.	Silicon Valley Bank	Snapdeal
45.	Delivery Hero	ColorChip	Foundation Capital	Silver Lake Ptrs
46.	Spotify	SpotHero	Mayfield Fund	Wefunder
47.	Univ. Studios Jp.	Path	Kima Ventures	Imagine K12
48.	Infor	Optimizely	IDG Capital Partners	Rocket Internet
49.	One97 Comm.	Beepi	Startupbootcamp	HIGHLINEvc
50.	Xiaomi	LeadGenius	CRV	One97 Comm.

C.4 Funding and network metrics distribution differences for Country, investor type and Economic category

Figure C.2: Comparison between Indegree and Funding distributions for Nationality.

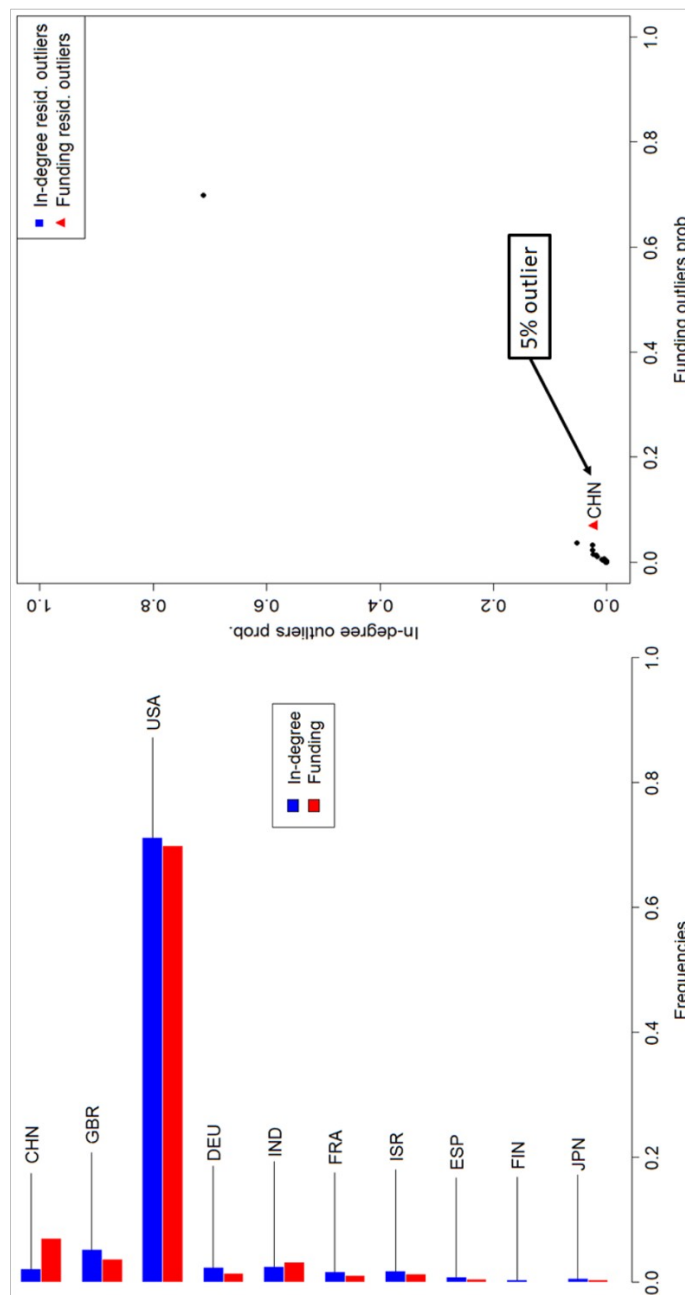


Figure C.3: Comparison between Indegree and Funding distributions for Investor Type.

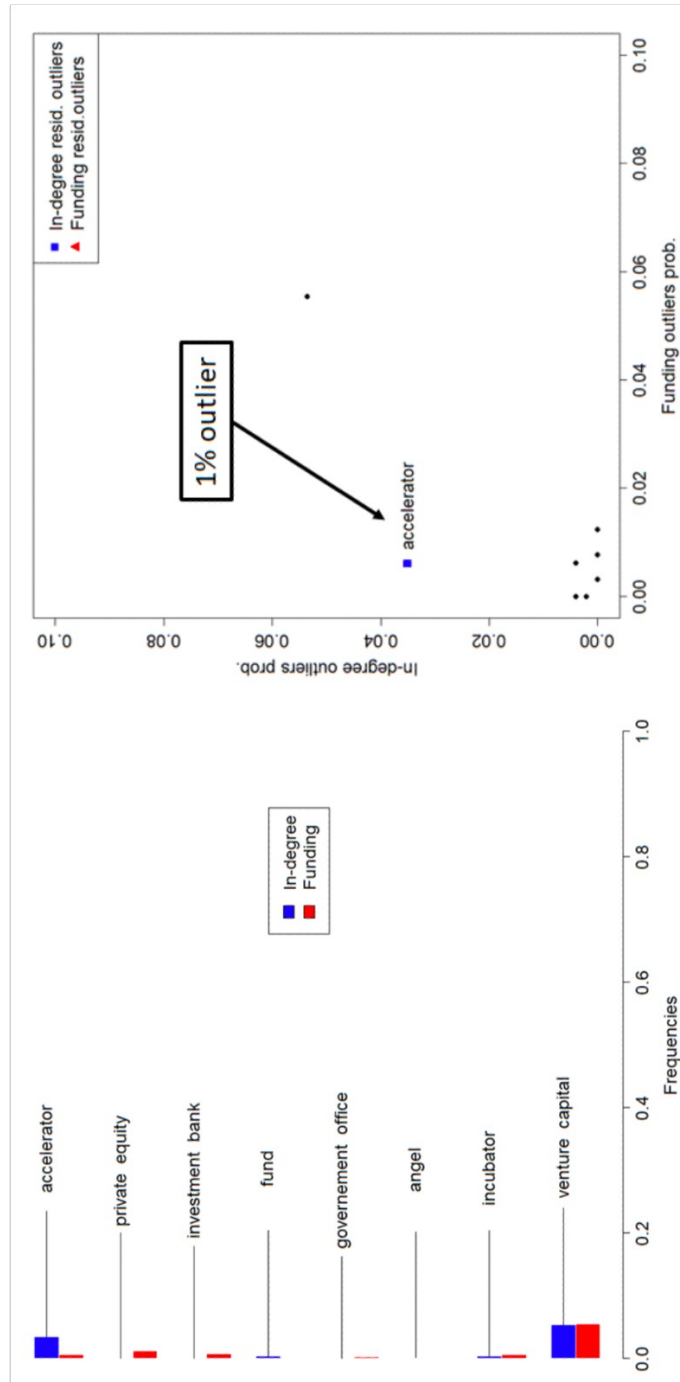


Figure C.4: Comparison between Indegree and Funding distributions for Economic Category.

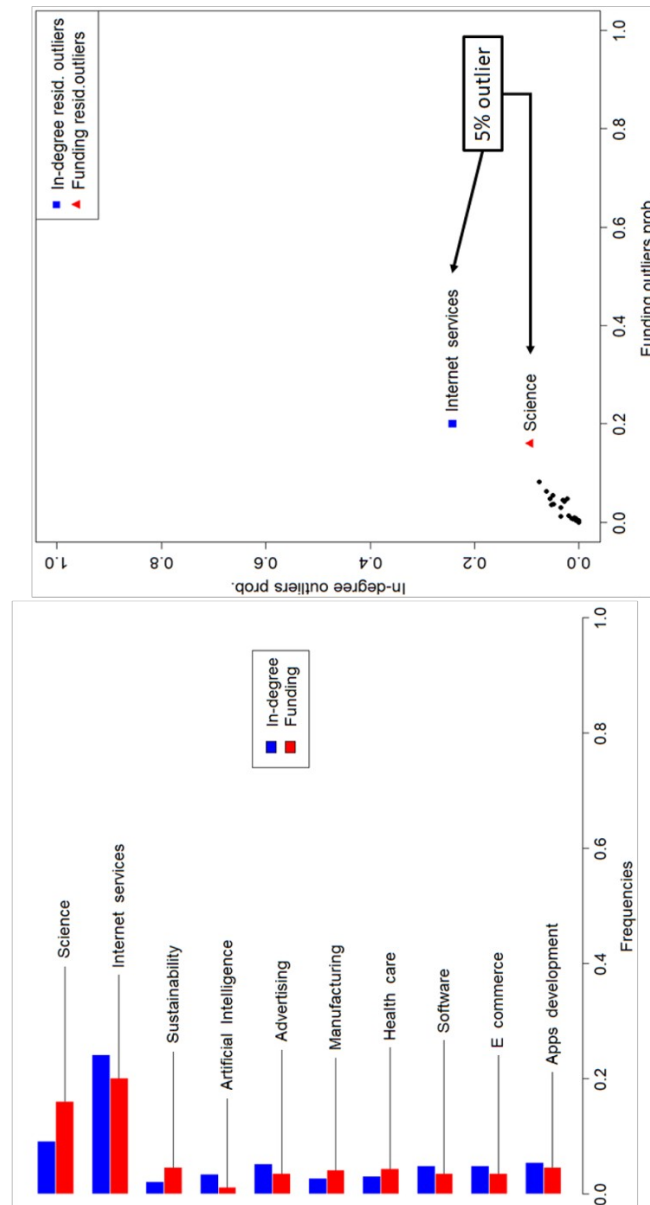


Figure C.5: Comparison between Outdegree and Funding distributions for Nationality.

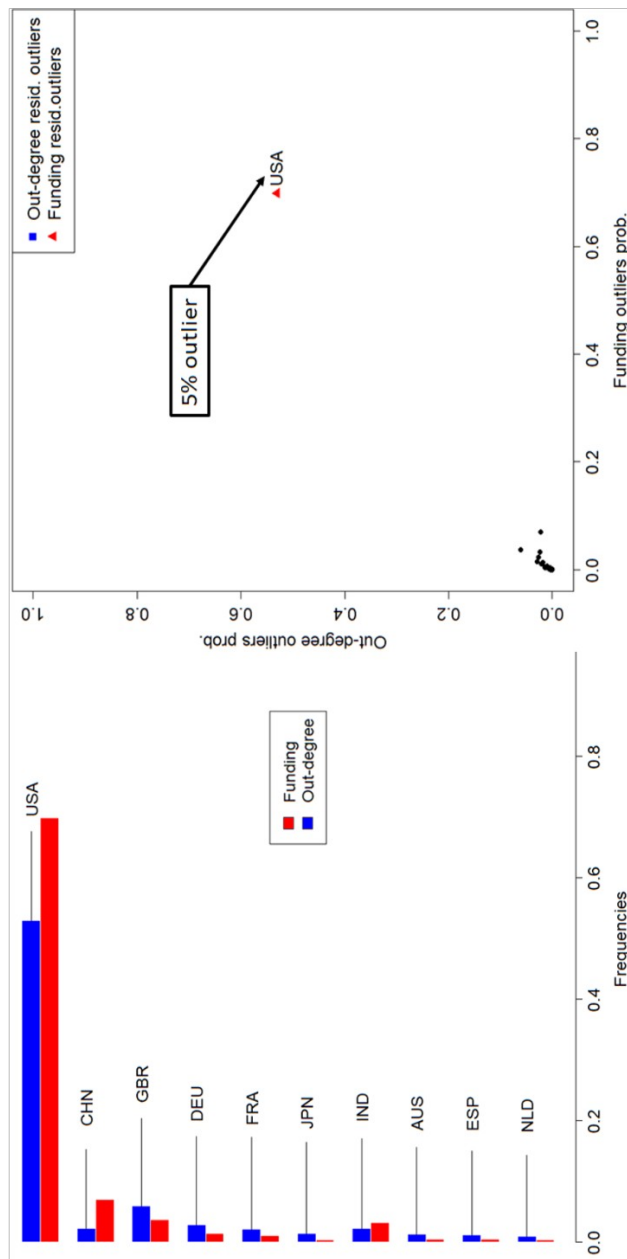


Figure C.6: Comparison between Outdegree and Funding distributions for Investor Type.

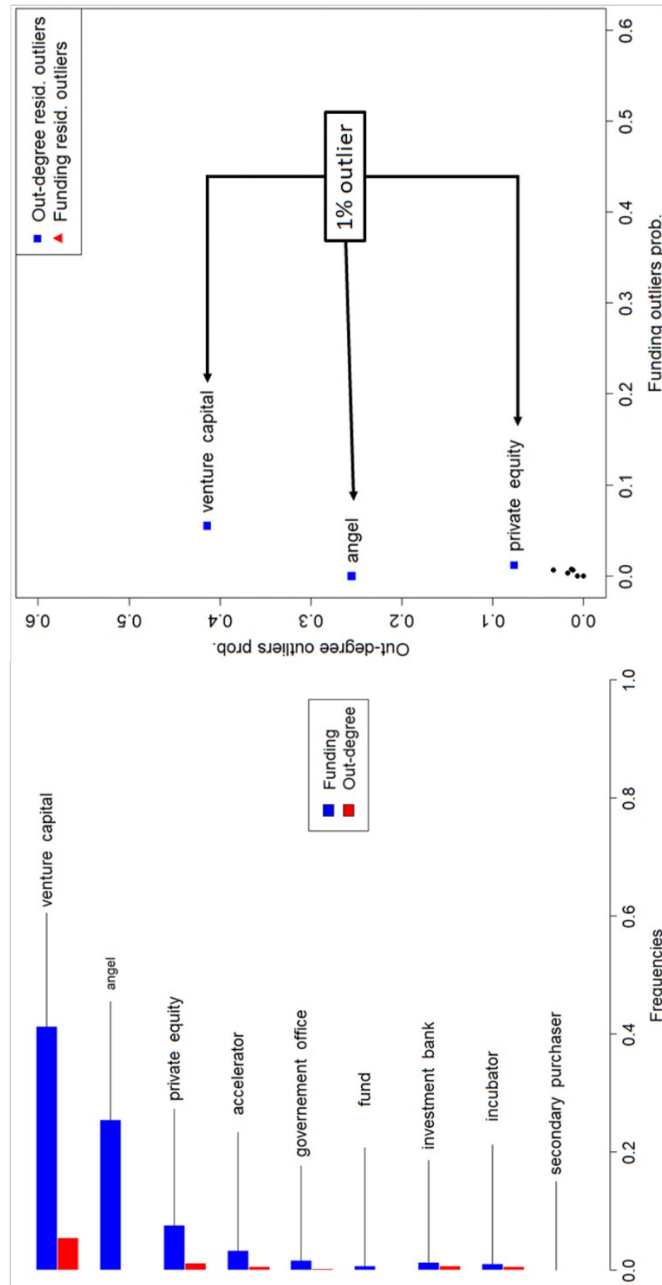


Figure C.7: Comparison between Outdegree and Funding distributions for Economic Category.

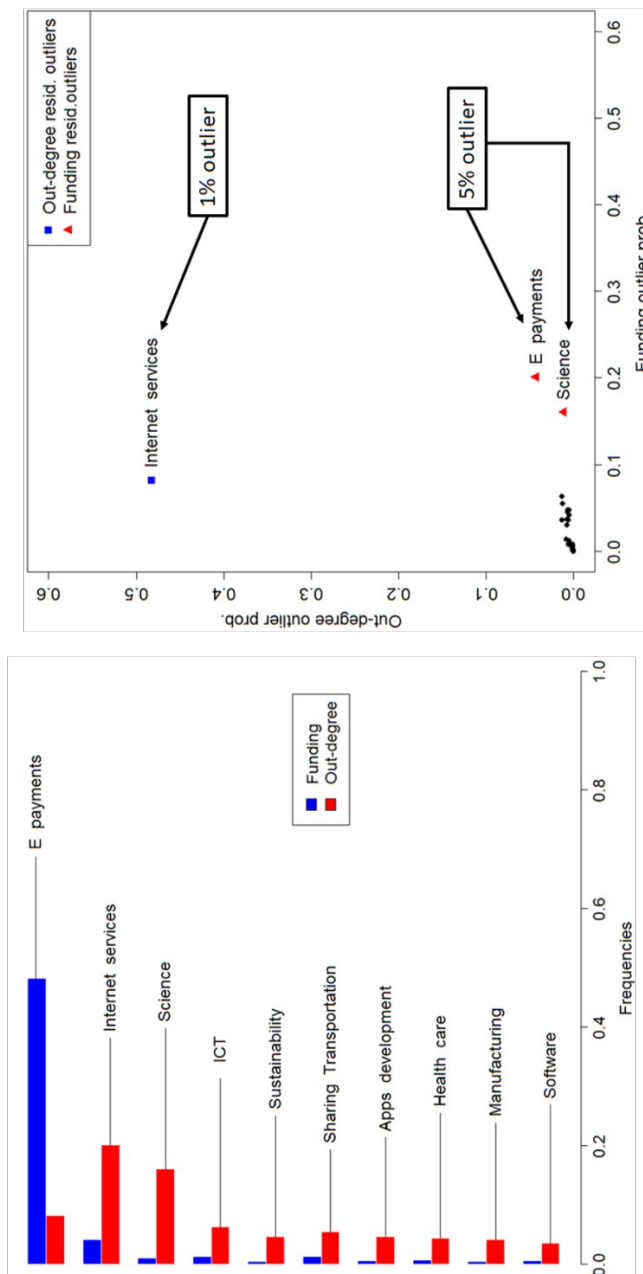


Figure C.8: Comparison between Betweenness and Funding distributions for Nationality.

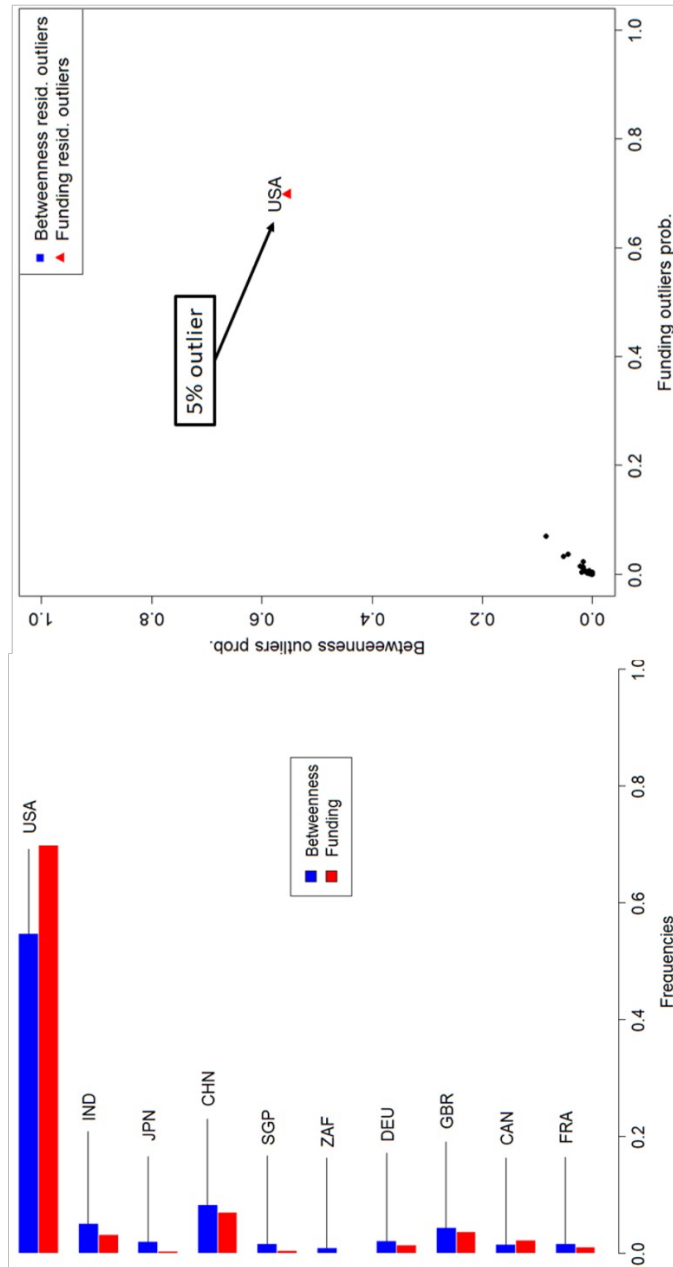


Figure C.9: Comparison between Betweenness and Funding distributions for Investor Type.

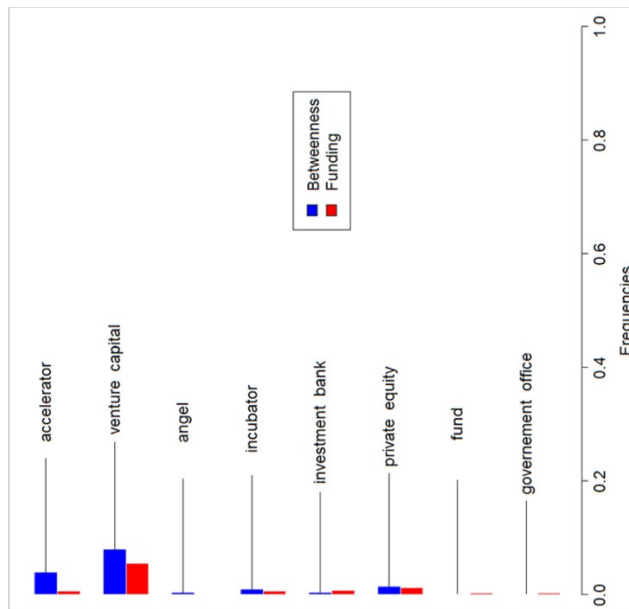
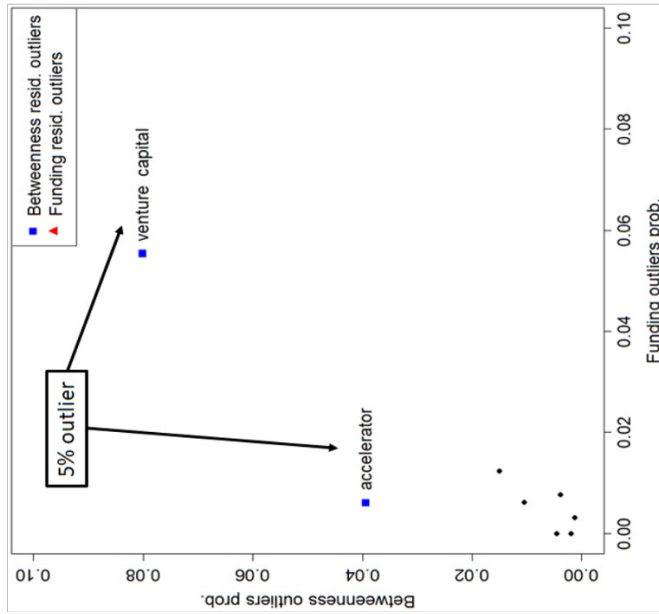
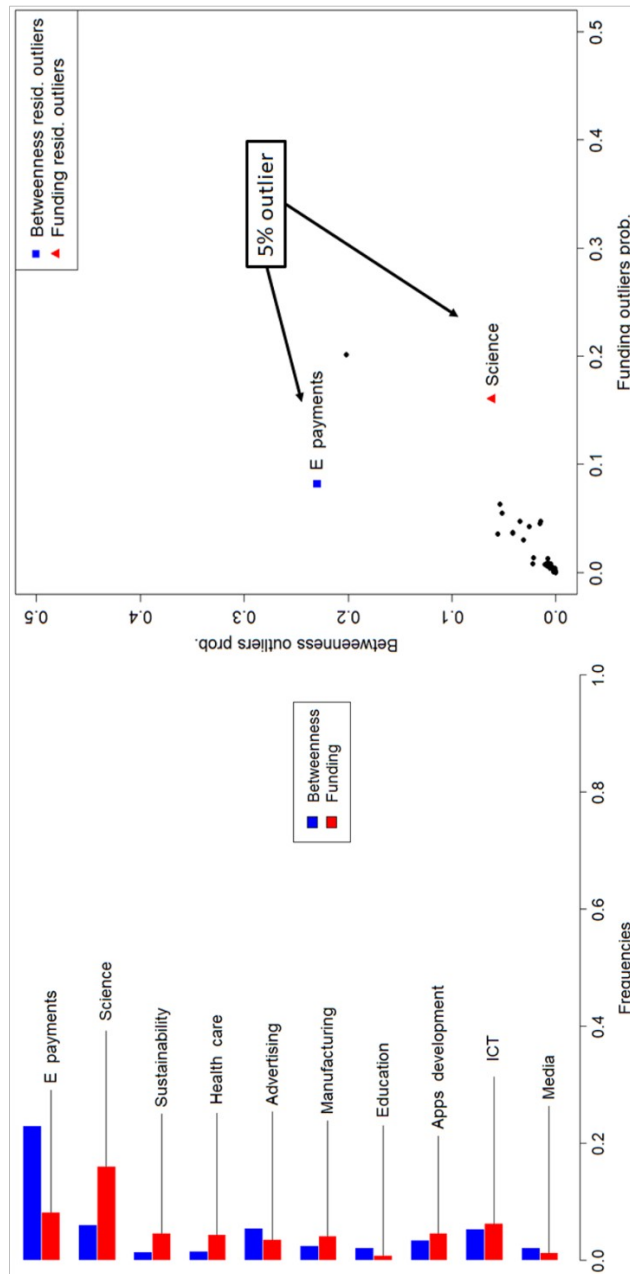


Figure C.10: Comparison between Betweenness and Funding distributions for Economic Category.



C.5 Statistical analyses

By definition, an outlier is an observation exceeding a distance of $1.5 \times IQR(X)$ from the first and third quartiles of its distribution X , where X = funding, indegree, outdegree and betweenness. Funding and centrality measures have long-tail distributions with large skewness and this feature may entail an overestimation of outliers. To tackle this issue, the experimental distributions have been bootstrapped ten thousands of times, estimating each time the left and right outlier thresholds and, finally, these results have been averaged. Accordingly, only observations exceeding these robust averaged left and right thresholds have been identified as outliers.

All statistical tests performed in this work are non-parametric Kolmogorov-Smirnov tests. Calculated p-values were corrected according to multiple hypothesis testing with Bonferroni correction.

Finally, to evaluate the effect of centrality measures on funding outliers, centrality measures have been considered for each year, from 2000 to 2017 and assigned the label 1 to those firms resulting funding outliers in the future 1, 2, . . . , 9 years. Thus 9 distinct datasets have been obtained, $d = 1, \dots, 9$ for each one, 100 10-fold cross-validation analyses have been performed in order to determine the model accuracy.

The findings presented in this work exploit the informative content provided by aggregate funds collected by each firm until 2017. Accordingly, it is possible to take into account:

- the information deriving from the overall temporal series of collected funds and exploiting it to obtain an accurate model of success;
- the economic interplay established over time and the bonds which therefore shape the network structure;

Considering funds collected over a long temporal range makes the aggregate network less sensitive to statistical fluctuations. Aggregating funds and therefore connections weakens the weight of each year with respect of the whole time series; the longer the series, the weaker the importance of each year. Therefore, aggregating information can be useful to explore global trends and strengthen the model's robustness.

Bibliography

- [1] Rabeh Morrar, Husam Arman, and Saeed Mousa. "The fourth industrial revolution (Industry 4.0): A social innovation perspective". In: *Technology innovation management review* 7.11 (2017), pp. 12–20.
- [2] Rainer Alt, Roman Beck and Martin T Smits. *FinTech and the transformation of the financial industry*. 2018.
- [3] Fábio Lotti Oliva et al. "Risks and critical success factors in the internationalization of born global startups of industry 4.0: A social, environmental, economic, and institutional analysis". In: *Technological Forecasting and Social Change* 175 (2022).
- [4] Zoltan J Acs and David B Audretsch. *Handbook of entrepreneurship research: An interdisciplinary survey and introduction*. Springer, 2010.
- [5] Steven J Davis et al. "Business volatility, job destruction and unemployment". In: *American Economic Journal: Macroeconomics* 2.2 (2010), pp. 259–287.
- [6] Giovani Da Silveira, Denis Borenstein, and Flavio S Fogliatto. "Mass customization: Literature review and research directions". In: *International journal of production economics* 72.1 (2001), pp. 1–13.
- [7] Zheng Xiang et al. "What can big data and text analytics tell us about hotel guest experience and satisfaction?" In: *International journal of hospitality management* 44 (2015), pp. 120–130.
- [8] Katerina Berezina et al. "Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews". In: *Journal of Hospitality Marketing & Management* 25.1 (2016), pp. 1–24.
- [9] Norman Au, Rob Law, and Dimitrios Buhalis. "The impact of culture on eComplaints: Evidence from Chinese consumers in hospitality organisations". In: *Information and communication technologies in tourism 2010*. Springer, 2010, pp. 285–296.
- [10] Robert Philip Weber. *Basic content analysis*. Vol. 49. Sage, 1990.
- [11] Zheng Xiang et al. "A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism". In: *Tourism Management* 58 (2017), pp. 51–65.
- [12] Yabing Zhao, Xun Xu, and Mingshu Wang. "Predicting overall customer satisfaction Big data evidence from hotel online textual reviews". In: *International Journal of Hospitality Management* 76 (2019) pp. 111–121.

- [13] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).
- [14] Mamad Mohamed. "Challenges and benefits of industry 4.0: An overview". In: *International Journal of Supply and Operations Management* 5.3 (2018), pp. 256-265.
- [15] Claudio Vitari and Elisabetta Raguseo. "Big data analytics business value and firm performance: Linking with environmental context". In: *International Journal of Production Research* 58.18 (2020), pp. 5456-5476.
- [16] Elisabetta Raguseo and Claudio Vitari. "Investments in big data analytics and firm performance: An empirical investigation of direct and mediating effects". In: *International Journal of Production Research* 56.15 (2018), pp. 5206-5221.
- [17] Klaus Schwab. *The fourth industrial revolution*. Currency, 2017.
- [18] Judith A Chevalier and Dina Mayzlin. "The effect of word of mouth on sales". In: *Journal of marketing research* 43.3 (2006), pp. 345-354.
- [19] Russell S Winer. "New communications approaches in marketing: Issues and research directions". In: *Journal of interactive marketing* 23.2 (2009), pp. 108-117.
- [20] Qiang Ye et al. "The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings". In: *Computers in Human behavior* 27.2 (2011), pp. 634-639.
- [21] Sangwon Park and Juan L Nicola. "Asymmetric effects of online consumer reviews". In: *Annals of Tourism Research* 50 (2015), pp. 67-83.
- [22] Yang Yang, Sangwon Park, and Xingbao Hu. "Electronic word of mouth and hotel performance: A meta-analysis". In: *Tourism management* 67 (2018), pp. 248-260.
- [23] Beverley A Sparks and Victoria Browning. "The impact of online reviews on hotel booking intentions and perception of trust". In: *Tourism management* 32.6 (2011), pp. 1310-1323.
- [24] SL Toral, MR Martínez-Torres, and MR Gonzalez-Rodriguez. "Identification of the unique attributes of tourist destinations from online reviews". In: *Journal of Travel Research* 57.7 (2018), pp. 908-919.
- [25] Aurelio G Mauri and Roberta Minazzi. "Web reviews influence on expectations and purchasing intentions of potential customers". In: *International journal of hospitality management* 34 (2013), pp. 99-107.
- [26] Peter O'connor. "Managing a hotel's image on TripAdvisor". In: *Journal of hospitality marketing & management* 19.7 (2010), pp. 754-772.
- [27] Hawoong Jeong et al. "Lethality and centrality in protein networks". In: *Nature* 411.6833 (2001), pp. 41-42.
- [28] Juan Camacho, Roger Guimerà, and Luís A Nunes Amaral. "Robust patterns in food web structure". In: *Physical Review Letters* 88.22 (2002), p. 228102.

- [29] Andrea Capocci et al. "Growing dynamics of internet providers" In: *Physical Review E* 64.3 (2001), p. 035105.
- [30] Alexei Vázquez, Romualdo Pastor-Satorras, and Alessandro Vespignani. "Large-scale topological and dynamic properties of the Internet" In: *Phys. Rev. E* 65 (6 June 2002), p. 066130.
- [31] Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. "Dynamic and correlation properties of the internet" *Physical review letters* 87.25 (2001), p. 258701.
- [32] Mark EJ Newman and Juyong Park. "Why social networks are different from other types of networks" In: *Physical review E* 68.3 (2003), p. 036122.
- [33] Mark EJ Newman. "Assortative mixing in networks". In: *Physical review letters* 89.20 (2002), p. 208701.
- [34] Paolo Giudici and Gianluca Passerone. "Data mining of association structures to model consumer behaviour". In: *Computational Statistics & Data Analysis* 38.4 (2002), pp. 533-541.
- [35] Jae Kyeong Kim et al. "Detecting the change of customer behavior based on decision tree analysis". In: *Expert Systems* 22.4 (2005), pp. 193-205.
- [36] Peter C Verhoef and Bas Donkers. "Predicting customer potential value an application in the insurance industry" *Decision support systems* 32.2 (2001), pp. 189-199.
- [37] Ning Sun et al. "iCARE: A framework for big data-based banking customer analytics". In: *IBM Journal of Research and Development* 58.5/6 (2014), pp. 4-1.
- [38] J Sophia Fu et al. "Two-stage modeling of customer choice preferences in engineering design using bipartite network analysis". In: *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Vol. 58127 American Society of Mechanical Engineers. 2017, V02AT03A039.
- [39] Mark Newman. *Networks*. Oxford university press, 2018.
- [40] Stefano Boccaletti et al. "Complex networks: Structure and dynamics". In: *Physics reports* 424.4-5 (2006), pp. 175-308.
- [41] Richard O Duda, Peter E Hart, and David G Stork. "Unsupervised learning and clustering". In: *Pattern classification* 2 (2001).
- [42] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.
- [43] Trevor Hastie et al. "Unsupervised learning". In: *The elements of statistical learning: Data mining, inference, and prediction* (2009).
- [44] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. "The k-means algorithm: A comprehensive survey and performance evaluation". In: *Electronics* 9.8 (2020).
- [45] Hae-Sang Park and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering". In: *Expert systems with applications* 36.2 (2009).

- [46] Fionn Murtagh and Pedro Contreras. "Algorithms for hierarchical clustering: an overview" In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97.
- [47] Santo Fortunato. "Community detection in graphs". *Physics reports* 486.3-5 (2010), pp. 75–174.
- [48] Punam Bedi and Chhavi Sharma. "Community detection in social networks" In: *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 6.3 (2016), pp. 115–135.
- [49] Andrea Lancichinetta and Santo Fortunato. "Community detection algorithms: a comparative analysis". *Physical review E* 80.5 (2009), p. 056117.
- [50] Jörg Reichardt and Stefan Bornholdt. "Statistical mechanics of community detection". In: *Physical review E* 74.1 (2006), p. 016110.
- [51] Vladimir Nasteski. "An overview of the supervised machine learning methods". In: *Horizons. b* 4 (2017), pp. 51–62.
- [52] Iqbal Muhammad and Zhu Yan. "SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY." In: *ICTACT Journal on Soft Computing* 5.3 (2015).
- [53] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. Mit Press, 2012.
- [54] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.
- [55] Noshir Contractor, Peter Monge, and Paul M Leonardi. "Network Theory | multidimensional networks and the dynamics of sociomateriality: bringing technology inside the network". *International Journal of Communication* 5 (2011), p. 39.
- [56] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of modern physics* 74.1 (2002), p. 47.
- [57] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *Science* 286.5439 (1999), pp. 509–512.
- [58] Tim S Evans and Bingsheng Chen. "Linking the network centrality measures closeness and degree". In: *Communications Physics* 5.1 (2022), p. 172.
- [59] Mark EJ Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical review E* 69.2 (2004), p. 026113.
- [60] Ahmad F Al Musawi, Satyaki Roy, and Preetam Ghosh. "Identifying accurate link predictors based on assortativity of complex networks". In: *Scientific Reports* 12.1 (2022), p. 18107.
- [61] Ulrik Brandes et al. "On finding graph clusterings with maximum modularity". In: *Graph-Theoretic Concepts in Computer Science: 33rd International Workshop, WG 2007, Dornburg, Germany, June 21-23, 2007. Revised Papers* 33. Springer, 2007, pp. 121–132.
- [62] Leon Danon et al. "Comparing community structure identification". *Journal of statistical mechanics: Theory and experiment* 2005.09 (2005), P09008.

- [63] Roger Guimera and Luís A Nunes Amaral. "Functional cartography of complex metabolic networks". In: *nature* 433.7028 (2005), pp. 895–900.
- [64] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. "Modularity from fluctuations in random graphs and complex networks". *Physical Review E* 70.2 (2004), p. 025101.
- [65] Andres Medus, Guillermo Acuna, and Claudio Oscar Dorso. "Detection of community structures in networks via global optimization". In: *Physica A: Statistical Mechanics and its Applications* 358.2-4 (2005) pp. 593–604.
- [66] Shuzhuo Liet et al. "A genetic algorithm with local search strategy for improved detection of community structure". In: *Complexity* 15.4 (2010), pp. 53–60.
- [67] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. "Finding community structure in very large networks". *Physical review E* 70.6 (2004), p. 066111.
- [68] Ken Wakita and Toshiyuki Tsurumi. "Finding community structure in mega-scale social networks". In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 1275–1276.
- [69] Suresh Shrivastava et al. "A review on credit card fraud detection using machine learning". In: *International Journal of Scientific & technology research* 8.10 (2019), pp. 1217–1220.
- [70] Vaishnavi Nath Dornadula and Sa Geetha. "Credit card fraud detection using machine learning algorithms". *Procedia computer science* 165 (2019), pp. 631–641.
- [71] Jan Salomon Cramer. "The early origins of the logit model". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 35.4 (2004) pp. 613–626.
- [72] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.
- [73] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.
- [74] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.
- [75] Carl Kingsford and Steven L Salzberg. "What are decision trees?". *Nature biotechnology* 26.9 (2008), pp. 1011–1013.
- [76] Tin Kam Ho. "A data complexity analysis of comparative advantages of decision forest construction". *Pattern Analysis & Applications* 5 (2002), pp. 102–112.
- [77] Weilun Wang, Goutam Chakraborty, and Basabi Chakraborty. "Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm". In: *Applied Sciences* 11.1 (2020), p. 202.
- [78] SL Ting, WH Ip, Albert HC Tsang, et al. "Is Naive Bayes a good classifier for document classification?". *International Journal of Software Engineering and Its Applications* 5.3 (2011), pp. 37–46.

- [79] Felix Abramovich, Vadim Grinshtein, and Tomer Levy. "Multiclass classification by sparse multinomial logistic regression". In: *IEEE Transactions on Information Theory* 67.7 (2021), pp. 4637–4646.
- [80] Angshuman Paul et al. "Improved random forest for classification". *IEEE Transactions on Image Processing* 27.8 (2018), pp. 4012–4024.
- [81] Shan Suthaharan and Shan Suthaharan. "Support vector machine". *Machine learning models and algorithms for big data classification: thinking with examples for effective learning* (2016), pp. 207–235.
- [82] Chaudhary Jashubhai Rameshbhai and Joy Paulose. "Opinion mining on newspaper headlines using SVM and NLP". In: *International journal of electrical and computer engineering (IJECE)* 9.3 (2019), pp. 2152–2163.
- [83] Francesco De Nicolò et al. "The verbalization of numbers: An explainable framework for tourism online reviews". In: *International Journal of Engineering Business Management* 15 (2023), p. 18479790231151913.
- [84] Stavros P Adam et al. "No free lunch theorem: A review". In: *Approximation and optimization: Algorithms, complexity and applications* (2019), pp. 57–82.
- [85] A Tharwat. *Classification assessment methods*. *Appl Comput Inform* 17 (1): 168–192. 2021.
- [86] André M Carrington et al. "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2022), pp. 329–341.
- [87] John Muschelli III. "ROC and AUC with a binary predictor: a potentially misleading metric". In: *Journal of classification* 37.3 (2020), pp. 696–708.
- [88] Khyati Chaudhary, Jyoti Yadav, and Bhawna Mallick. "A review of fraud detection techniques: Credit card". In: *International Journal of Computer Applications* 45.1 (2012), pp. 39–44.
- [89] Nicola Amoroso et al. "Deep learning reveals Alzheimer's disease onset in MCI subjects: results from an international challenge". In: *Journal of neuroscience methods* 302 (2018), pp. 3–9.
- [90] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". *Artificial intelligence* 1.5 (2019), pp. 206–215.
- [91] Luke Merrick and Ankur Taly. "The explanation game: Explaining machine learning models using shapley values". *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings* 4. Springer. 2020, pp. 17–38.
- [92] Lloyd S Shapley et al. "A value for n-person games". In: (1953).
- [93] Michael Maschler, Shmuel Zamir, and Eilon Solan. *Game theory*. Cambridge University Press, 2020.
- [94] Tavneet Suri et al. "Paths to success: The relationship between human development and economic growth". In: *World Development* 39.4 (2011), pp. 506–522.

- [95] Richard Florida. "The creative class and economic development". In: *Economic development quarterly* 28.3 (2014), pp. 196–205.
- [96] Andrea Tacchella et al. "A new metrics for countries' fitness and products' complexity". In: *Scientific reports* 2.1 (2012), p. 723.
- [97] Penny Mealy, J Doyne Farmer, and Alexander Teytelboym. "Interpreting economic complexity". In: *Science advances* 5.1 (2019), eaau1705.
- [98] Yian Yin et al. "Quantifying the dynamics of failure across science, startups and security". In: *Nature* 575.7781 (2019), pp. 190–194.
- [99] Moreno Bonaventura et al. "Predicting success in the worldwide start-up network". In: *Scientific reports* 10.1 (2020), p. 345.
- [100] Sameh Al-Natour and Ozgur Turekci. "A comparative assessment of sentiment analysis and star ratings for consumer reviews". *International Journal of Information Management* 54 (2020), p. 102132.
- [101] Karen Robson et al. "Making sense of online consumer reviews: A methodology". In: *International Journal of Market Research* 55.4 (2013), pp. 521–537.
- [102] Sung Ho Ha, SY Bae, and Lee Kyeong Son. "Impact of online consumer reviews on product sales: Quantitative analysis of the source effect". In: *Applied Mathematics and Information Sciences* 9.2L (2015), pp. 373–387.
- [103] Miriam Alzate, Marta Arce-Urriza, and Javier Cebollada. "Mining the text of online consumer reviews to analyze brand image and brand positioning". In: *Journal of Retailing and Consumer Services* 67 (2022), p. 102989.
- [104] Albert-László Barabási. *The formula: The universal laws of success*. Hachette UK, 2018.
- [105] Nicola Amoroso et al. "Economic interplay forecasting business success". In: *Complexity* 2021 (2021), pp. 1–12.
- [106] Loredana Bellantuono et al. "An equity-oriented rethink of global rankings with complex networks mapping development". In: *Scientific Reports* 10.1 (2020), p. 18046.
- [107] An Zeng et al. "The science of science: From the perspective of complex systems". In: *Physics reports* 714 (2017), pp. 1–73.
- [108] Stephen A Gallo, Joanne H Sullivan, and Scott R Glisson. "The influence of peer reviewer expertise on the evaluation of research funding applications". In: *PloS one* 11.10 (2016), e0165147.
- [109] Bruno Gonçalves et al. "Exploring team passing networks and player movement dynamics in youth association football". *PloS one* 12.1 (2017), e0171156.
- [110] Péter Érdi et al. "Prediction of emerging technologies based on analysis of the US patent citation network". In: *Scientometrics* 95 (2013), pp. 225–242.
- [111] Debora Valentina Malito, Gaby Umbach, and Nehal Bhuta. *The Palgrave handbook of indicators in global governance*. Springer, 2018.

- [112] Alexander Cooley and Jack Snyder. *Ranking the world*. Cambridge University Press, 2015.
- [113] Jerry Muller. *The tyranny of metrics*. Princeton University Press, 2018.
- [114] Péter Érdi. *Ranking: The unwritten rules of the social game we all play*. Oxford University Press, 2019.
- [115] Jennifer Clark. *Uneven innovation: The work of smart cities*. Columbia University Press, 2020.
- [116] Brad Feld. *Startup communities: Building an entrepreneurial ecosystem in your city*. John Wiley & Sons, 2020.
- [117] Stefania Fiorentino. *Startup cities: Why only a few cities dominate the global startup scene and what the rest should do about it: by Peter S. Cohan, Apress, Marlborough, 2018. 271 pp., US 29.99(pbk), ISBN 13 : 978-1-4842-3392-4, https : //www.springer.com/la/book/9781484233924. 2020.*
- [118] *StartupBlink Startup Ecosystem Rankings 2017*. <https://www.startupblink.com/startups>. Accessed: 2 September 2022.
- [119] *StartupGenome Global Startup Ecosystem Report 2016*. <https://www.startupgenome.com/all-reports>. Accessed: 2 September 2022.
- [120] *Crunchbase: Discover innovative companies and the people behind them*. <https://www.crunchbase.com>. Accessed: 2 September 2022.
- [121] Amy N Langville and Carl Meyer. *Who's # 1? The science of rating and ranking*. Princeton University Press, 2012.
- [122] Francesco De Nicolò et al. "Territorial Development as an Innovation Driver: A Complex Network Approach". In: *Applied Sciences* 12.18 (2022), p. 9069.
- [123] Elena Esposito and David Stark. "What's Observed in a Rating? Rankings as Orientation in the Face of Uncertainty". In: *Theory, Culture & Society* 36.4 (2019), pp. 3-26.
- [124] Marta Kuc-Czarnecka, Samuele Lo Piano and Andrea Saltelli. "Quantitative storytelling in the making of composite indicators". *Social Indicators Research* 149.3 (2020), pp. 775-802.
- [125] Anshul Verma, Orazio Angelini and Tiziana Di Matteo. "A new set of cluster driven composite development indicators". In: *EPJ Data Science* 9.1 (2020), p. 8.
- [126] Bjørn Høyland, Karl Moene, and Fredrik Willumsen. "The tyranny of international index rankings". In: *Journal of Development economics* 97.1 (2012), pp. 1-14.
- [127] César A Hidalgo et al. "The product space conditions the development of nations". In: *Science* 317.5837 (2007), pp. 482-487.
- [128] *World Development Indicators - Databank*. <https://databank.worldbank.org/source/world-development-indicators>. Accessed 2 September 2022.
- [129] Attila Lajos Makai. "Startup Ecosystems Rankings". In: *Hungarian Statistical Review* 4 (2) (2021), pp. 70-94.

- [130] Riitta Katila, Eric L Chen, and Henning Piezunka. "All the right moves: How entrepreneurial firms compete effectively". In: *Strategic Entrepreneurship Journal* 6.2 (2012), pp. 116-132.
- [131] Susan Cohen. "What do accelerators do? Insights from incubators and angels". In: *Innovations: Technology, Governance, Globalization* 8.3 (2013), pp. 19-25.
- [132] *Doing Business Report 2019*. <https://archive.doingbusiness.org/>. Accessed: 2 September 2022.
- [133] Peter Witt. "Entrepreneurs, networks and the success of start-ups". *Entrepreneurship & Regional Development* 16.5 (2004), pp. 391-412.
- [134] Agnes Dessyana and Benedicta Prihatin Dwi Riyanti. "The influence of innovation and entrepreneurial self-efficacy to digital startup success". In: *International research journal of business studies* 10.1 (2017), pp. 57-68.
- [135] Aidin Salamzadeh. *Start-up boom in an emerging market: A niche market approach*. Springer, 2018.
- [136] Jean-Michel Dalle, Matthijs Den Besten, and Carlo Menon. "Using Crunchbase for economic and managerial research". In: (2017).
- [137] Oliver Alexy et al. "The social capital of venture capitalists and its impact on the funding of start-up firms". In: *ERIM Report Series Reference No. ERS-2010-028-ORG* (2010).
- [138] Oliver T Alexy et al. "Social capital of venture capitalists and start-up funding". In: *Small Business Economics* 39 (2012), pp. 835-851.
- [139] Amulya Tata et al. "The psycholinguistics of entrepreneurship". In: *Journal of Business Venturing Insights* 7 (2017), pp. 38-44.
- [140] Anne LJ Ter Wal et al. "The best of both worlds: The benefits of open-specialized and closed-diverse syndication networks for new ventures' success". In: *Administrative science quarterly* 61.3 (2016), pp. 393-432.
- [141] Jessica Santana, Raine Hoover, and Meera Vengadasubbu. "Investor commitment to serial entrepreneurs: A multilayer network analysis". In: *Social Networks* 48 (2017), pp. 256-269.
- [142] Rossella Pozzi, Tommaso Rossi, and Raffaele Secchi. "Industry 4.0 technologies critical success factors for implementation and improvements in manufacturing companies". *Production Planning & Control* 34.2 (2023), pp. 139-158.
- [143] Uglješa Stankov and Ulrike Gretzel. "Tourism 4.0 technologies and tourist experiences: a human-centered design perspective". In: *Information Technology & Tourism* 22.3 (2020), pp. 477-488.
- [144] Gulnora Kalandarovna Abdurakhmanova et al. "TOURISM 4.0: OPPORTUNITIES FOR APPLYING INDUSTRY 4.0 TECHNOLOGIES IN TOURISM". In: *Proceedings of the 6th International Conference on Future Networks & Distributed Systems*. 2022, pp. 33-38.
- [145] Philippe Duverger. "Curvilinear effects of user-generated content on hotels' market share: a dynamic panel-data analysis". In: *Journal of Travel Research* 52.4 (2013), pp. 465-478.

- [146] ZhiweiLiu and Sangwon Park.“What makes a useful online review? Implication for travel product websites”. In: *Tourism management* 47 (2015), pp. 140–151.
- [147] Julian K Ayeh, Norman Au, and Rob Law. “Do we believe in TripAdvisor? Examining credibility perceptions and online travel site use toward using user-generated content”. In: *Journal of Travel Research* 52.4 (2013), pp. 437–452.
- [148] Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. “Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content”. In: *Marketing Science* 31.3 (2012), pp. 493–520.
- [149] Luis V Casalo et al. “Avoiding the dark side of positive online consumer reviews: Enhancing reviews’ usefulness for high risk-averse travelers”. In: *Journal of Business Research* 68.9 (2015), pp. 1829–1835.
- [150] Yue Pan and Jason Q Zhang. “Born unequal: a study of the helpfulness of user-generated product reviews”. In: *Journal of Retailing* 87.4 (2011), pp. 598–612.
- [151] Senga Briggs, Jean Sutherland, and Siobhan Drummond. “Are hotels serving quality? An exploratory study of service quality in the Scottish hotel sector”. In: *Tourism management* 28.4 (2007), pp. 1006–1019.
- [152] Andrei P Kirilenko et al. “Automated sentiment analysis in tourism: Comparison of approaches”. In: *Journal of Travel Research* 57.8 (2018), pp. 1012–1025.
- [153] Linchi Kwok. “Exploratory-triangulation design in mixed methods studies: A case of examining graduating seniors who meet hospitality recruiters’ selection criteria”. In: *Tourism and Hospitality Research* 12.3 (2012), pp. 125–138.
- [154] Weilin Lu and Svetlana Stepchenkova. “Ecotourism experiences reported online: Classification of satisfaction attributes”. In: *Tourism management* 33.3 (2012), pp. 702–712.
- [155] Praphula Kumar Jain et al. “Consumer recommendation prediction in online reviews using Cuckoo optimized machine learning models”. *Computers and Electrical Engineering* 95 (2021), p. 107397.
- [156] Barkha Bansal and Sangeet Srivastava. “Hybrid attribute based sentiment classification of online reviews for consumer intelligence”. *Applied Intelligence* 49.1 (2019), pp. 137–149.
- [157] Nikhil Kumar Singh, Deepak Singh Tomar, and Arun Kumar Sangaiah. “Sentiment analysis, review and comparative analysis over social media”. In: *Journal of Ambient Intelligence and Humanized Computing* 11 (2020), pp. 97–117.
- [158] Yi-Chun Ho, Junjie Wu, and Yong Tan. “Disconfirmation effect on online rating behavior: A structural model”. In: *Information Systems Research* 28.3 (2017), pp. 626–642.
- [159] Truc H Le et al. “Proposing a systematic approach for integrating traditional research methods into machine learning in text analytics in tourism and hospitality”. In: *Current Issues in Tourism* 24.12 (2021), pp. 1640–1655.

- [160] Tianxiang Zheng et al. "Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach". In: *International Journal of Hospitality Management* 92 (2021), p. 102658.
- [161] *TripAdvisor*. <http://tripadvisor.it>. Accessed: 20 September 2023.
- [162] K.R. Chowdhary. *Fundamentals of Artificial Intelligence*. Feb. 2020.
- [163] Chris Forman, Anindya Ghose, and Batia Wiesenfeld. "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets". In: *Information systems research* 19.3 (2008), pp. 291-313.
- [164] Serhad Sarica and Jianxi Luo. "Stopwords in technical language processing". In: *Plos one* 16.8 (2021), e0254937.
- [165] Ronen Feldman. "Techniques and applications for sentiment analysis". In: *Communications of the ACM* 56.4 (2013), pp. 82-89.
- [166] Mingqing Hu and Bing Liu. "Mining and summarizing customer reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 168-177.
- [167] Vincent A Traag and Jeroen Bruggeman. "Community detection in networks with positive and negative links". In: *Physical Review E* 80.3 (2009), p. 036115.
- [168] Vincent A Traag, Ludo Waltman and Nees Jan Van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific reports* 9.1 (2019), p. 5233.
- [169] Fa-Yueh Wu. "The potts model". In: *Reviews of modern physics* 54.1 (1982), p. 235.
- [170] Vincent A Traag. "Faster unfolding of communities: Speeding up the Louvain algorithm". In: *Physical Review E* 92.3 (2015), p. 032801.
- [171] Gergely Palla et al. "Hierarchical networks of scientific journals". In: *Palgrave Communications* 1.1 (2015), pp. 1-9.
- [172] Alex Arenas et al. "Community analysis in social networks". In: *The European Physical Journal B* 38 (2004), pp. 373-380.
- [173] Onno Hoffmeister. "Development status as a measure of development". In: *Statistical Journal of the IAOS* 36.4 (2020), pp. 1095-1128.
- [174] *The International Standard for country codes and codes for their subdivisions - ISO 3166 country codes*. <https://databank.worldbank.org/source/world-development-indicators>. Accessed: 2 September 2022.
- [175] *How does the World Bank classify countries?* <https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries.html>. Accessed: 2 September 2022.
- [176] Linton C Freeman et al. "Centrality in social networks: Conceptual clarification". In: *Social network: critical concepts in sociology*. Londres: Routledge 1 (2002), pp. 238-263.
- [177] Noah E Friedkin. "Theoretical foundations for centrality measures". *American journal of Sociology* 96.6 (1991), pp. 1478-1504.

- [178] Sea Jin Chang. "Venture capital financing, strategic alliances and the initial public offerings of Internet startups". In: *Journal of Business Venturing* 19.5 (2004), pp. 721-741.
- [179] Josh Lerner et al. "The globalization of angel investments: Evidence across countries". In: *Journal of Financial Economics* 127.1 (2018), pp. 1-20.
- [180] Vijith M Nair and Dileep G Menon. "Fin Tech firms-A new challenge to Traditional Banks: A Review". In: *International Journal of Applied Business and Economic Research* 15. Special Issue (2017), pp. 173-184.
- [181] David A Kenny. *Statistics for the social and behavioral sciences*. Brown, 1987.
- [182] Akiko Aizawa. "An information-theoretic perspective of tf-idf measures". In: *Information Processing & Management* 39.1 (2003), pp. 45-65.
- [183] Shahzad Qaiser and Ramsha Ali. "Text mining: use of TF-IDF to examine the relevance of words to documents". In: *International Journal of Computer Applications* 181.1 (2018), pp. 25-29.
- [184] Vaishali Ganganwar. "An overview of classification algorithms for imbalanced datasets". In: *International Journal of Emerging Technology and Advanced Engineering* 2.4 (2012), pp. 42-47.
- [185] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216-225.
- [186] José A Sáez, Julián Luengo, and Francisco Herrera. "Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure". In: *Neurocomputing* 176 (2016), pp. 26-35.
- [187] Xingquan Zhu and Xindong Wu. "Class noise vs. attribute noise: A quantitative study". In: *Artificial Intelligence review* 22 (2004), pp. 177-210.
- [188] Mário Antunes et al. "Knee/elbow point estimation through thresholding". In: *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*. IEEE. 2018, pp. 413-419.
- [189] Lior Rokach and Oded Maimon. "Clustering methods". In: *Data mining and knowledge discovery handbook* (2005), pp. 321-352.
- [190] Richard C Geibel and Meghana Manickam. "Comparison of selected startup ecosystems in Germany and in the USA Explorative analysis of the startup environments". In: *GSTF Journal on Business Review (GBR)* 4.3 (2016).
- [191] Itxaso del-Palacio and Dave Chapman. "United Kingdom: London's tech startup boom". In: *Global Clusters of Innovation*. Cheltenham, England: Edward Elgar Publishing, 2014.
- [192] *StartupBlink Startup Ecosystem Rankings 2019*. <https://www.startupblink.com/startups>. Accessed: 2 September 2022.

- [193] Shaker A Zahra and Niron Hashai. "The effect of MNEs' technology startup acquisitions on small open economies' entrepreneurial ecosystems". In: *Journal of International Business Policy* 5.3 (2022), pp. 277–295.
- [194] Zahra Nazari et al. "Evaluation of class noise impact on performance of machine learning algorithms". *IJCSNS Int. J. Comput. Sci. Netw. Secur* 18 (2018), p. 149.
- [195] Lusiana Citra Dewi Alvin Chandra, et al. "Social media web scraping using social media developers API and regex". *Procedia Computer Science* 157 (2019), pp. 444–449.
- [196] Nisha Rathee, Nikita Joshi, and Jaspreet Kaur. "Sentiment analysis using machine learning techniques on Python". *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE. 2018, pp. 779–785.
- [197] Muzaffer Can Iban and Alihsan Sekertekin. "Machine learning based wildfire susceptibility mapping using remotely sensed fire data and GIS: A case study of Adana and Mersin provinces, Turkey". In: *Ecological Informatics* 69 (2022), p. 101647.
- [198] Roberto Cilli et al. "Explainable artificial intelligence (XAI) detects wildfire occurrence in the Mediterranean countries of Southern Europe". In: *Scientific reports* 12.1 (2022), p. 16349.
- [199] Angela Lombardi et al. "Explainable deep learning for personalized age prediction with brain morphology". In: *Frontiers in neuroscience* 15 (2021), p. 578.
- [200] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. "Drug discovery with explainable artificial intelligence". In: *Nature Machine Intelligence* 2.10 (2020), pp. 573–584.
- [201] Siwei Lai et al. "Recurrent convolutional neural networks for text classification". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 29. 1. 2015.
- [202] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. "Recurrent neural networks for time series forecasting: Current status and future directions". *International Journal of Forecasting* 37.1 (2021), pp. 388–427.
- [203] Dominic Gorecky et al. "Human-machine-interaction in the industry 4.0 era". In: *2014 12th IEEE international conference on industrial informatics (INDIN)*. IEEE. 2014, pp. 289–294.
- [204] Katrin Scheibe, Kaja J Fietkiewicz, and Wolfgang G Stock. "Information behavior on live streaming services". *Journal of Information Science Theory and Practice* 4.2 (2016), pp. 6–20.
- [205] Deepjyoti Roy and Mala Dutta. "A systematic review and research perspective on recommender systems". *Journal of Big Data* 9.1 (2022), p. 59.
- [206] Meenakshi Sharma and Sandeep Mann. "A survey of recommender systems: approaches and limitations". *International journal of innovations in engineering and technology* 2.2 (2013), pp. 8–14.

- [207] Jie Zhou et al. "Graph neural networks: A review of methods and applications". In: *AI open* 1 (2020), pp. 57–81.
- [208] Zonghan Wu et al. "A comprehensive survey on graph neural networks". In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.
- [209] Wenqi Fan et al. "Graph neural networks for social recommendation". In: *The world wide web conference*. 2019, pp. 417–426.
- [210] Zhan Shi et al. "Smart factory in Industry 4.0". In: *Systems Research and Behavioral Science* 37.4 (2020), pp. 607–617.
- [211] Marcel Matthes et al. "Supplier sustainability assessment in the age of Industry 4.0-Insights from the electronics industry". In: *Cleaner logistics and supply chain* 4 (2022), p. 100038.
- [212] Jay Lee, Behrad Bagheri and Hung-An Kao. "Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics". In: *International proceeding of international conference on industrial informatics (INDIN)*. Citeseer. 2014, pp. 1–6.
- [213] Surajit Bag and Jan Harm Christiaan Pretorius. "Relationships between industry 4.0, sustainable manufacturing and circular economy: proposal of a research framework". In: *International Journal of Organizational Analysis* 30.4 (2022), pp. 864–898.
- [214] Amin Dehdarian and Christopher L Tucci. "A complex network approach for analyzing early evolution of smart grid innovations in Europe". In: *Applied Energy* 298 (2021), p. 117143.
- [215] Carlos Díaz-Santamaría and Jacques Bulchand-Giduffo. "Econometric estimation of the factors that influence startup success". In: *Sustainability* 13.4 (2021), p. 2242.
- [216] Mariapina Trunfio, Luca Petruzzellis, and Claudio Nigro. "Tour operators and alternative tourism in Italy: Exploiting niche markets to increase international competitiveness". In: *International Journal of Contemporary Hospitality Management* 18.5 (2006), pp. 426–438.
- [217] Noor Kamal Kaur, Usvir Kaur, and Dheerendra Singh. "K-Medoid clustering algorithm-a review". In: *Int. J. Comput. Appl. Technol.* (2014), pp. 42–45.
- [218] Ketan Rajshekhar Shahapure and Charles Nicholas. "Cluster quality analysis using silhouette score". In: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE. 2020, pp. 747–748.
- [219] Rena Nainggolan et al. "Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method". In: *Journal of Physics: Conference Series*. Vol. 1361.1. IOP Publishing. 2019, p. 012015.
- [220] Weksi Budiaji and Friedrich Leisch. "Simple K-medoids partitioning algorithm for mixed variable data". In: *Algorithms* 12.9 (2019), p. 177.