UNIVERSITÀ DI PISA
Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

Social Media Analytics and Open Source Intelligence: the role of Social Media in intelligence activities

Mariantonietta Noemi La Polla

Anno 2014

UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

Social Media Analytics and Open Source Intelligence: the role of social media in intelligence activities

Autore:

Mariantonietta Noemi La Polla _____

Relatore:

Prof. Francesco Marcelloni _____

Dott. Andrea Marchetti _____

Dott..Maurizio Tesconi _____

Anno 2014
SSD ING-INF/05

1

Sommario

Un numero sempre crescente di forze dell'ordine di tutto il mondo utilizza anche i social media per le proprie investigazioni. Il motivo principale è che le organizzazioni criminali utilizzano sempre più i social media per le loro attività di propaganda e arruolamento. Lo sfruttamento di dati provenienti dai social media per identificare e tracciare le attività delle organizzazioni criminali è alla base della *social media intelligence* (SOCMINT). L'approccio tradizionale per le indagini criminali, attualmente utilizzato in ambienti di *intelligence*, non può essere applicato però così com'è per l'analisi dei *social network*. Questo sopratutto perché ci sono vincoli dovuti al rispetto della *privacy* degli utenti che utilizzano questi nuovi strumenti di comunicazione. Considerata anche la tendenza, sempre più diffusa tra gli utenti, a nascondere le proprie informazioni agli utenti non amici , nei dataset da analizzare le impostazioni di *privacy* possono portare alla formazione di buchi neri. Si tratta della mancanza di alcune informazioni relative, ad esempio al profilo di un utente quali il numero di amici, che possono rilevarsi importanti ai fini investigativi. In questa tesi viene presentato un approccio per le analisi dei dati dai social media per superare questo problema, spostando l'attenzione su spazi di social media pubblici (*public-by-design*). L'utilizzo dei dati provenienti da questi spazi risolve i problemi relativi alla *privacy*, quelli etici e quelli legati ai limiti tecnici, buchi neri inclusi. Viene inizialmente descritto un *framework* che consenta di gestire una *pipeline* oper le analisi dei dati da social media. Il *framework* è modulare ed è stato progettato in modo da essere il più indipendente possibile dal social media considerato. Viene poi presentato un algoritmo che recupera il *grafo delle interazioni* tra gli utenti, a partire dai contenuti scambiati sugli spazi pubblici di Facebook. Questo grafo descrive come gli utenti interagiscono fra loro nel social media e può essere studiato con strumenti di *social media analytics*. Infine viene descritto un sistema di pesi

per il grafo delle interazioni. Questo sistema è stato studiato in modo da prendere in considerazione la natura differente delle numerose interazioni possibili. Per validare lo studio è stato utilizzato un gruppo reale di utenti Facebook. I membri del gruppo sono stati intervistati in modo da i) determinare il grafo di interazione reale; ii) avere indicazioni utili alla fase di assegnazione dei pesi. I risultati ottenuti mostrano che l'analisi delle interazioni consente di evidenziare relazioni tra utenti che sono utili ai fini della attività investigative. Il grafo delle interazioni fornisce, dunque, una visione più di completa di come gli utenti interagiscono fra loro. Inoltre, i risultati ottenuti confermano l'importanza degli utenti attivi di una rete.

2

Abstract

Nowadays a growing number of law enforcement agencies use social media even for criminal investigations. This because a growing number of criminal organizations daily use and rely on social media for their illegal activities of propaganda and criminals' enrollment. The exploitation of social media data to track the activities of criminal organizations is part of the social media intelligence (SOCMINT). The classical criminals analysis, performed in an intelligence environment, cannot be conducted in the same way as, for instance, social networks analysis. This is due to limitations imposed by privacy and ethical concerns. Privacy settings can cause black holes in the analyzed data because people tend more often to hide their information. As a result, the analysis cannot be complete and meaningful. This thesis tries to overcome this problem by moving the attention into public-by-design spaces in social media. Using data coming from public spaces solves the problems related to privacy, ethical and related technical limitations, including the black holes. First, a framework to easily manage social media pipeline is presented. The framework is modular and is designed in order to be as much independent as possible from the social media considered. The proposed algorithm retrieves from public entities in social media the interaction graph. This graph describes how users interact each other in social media and can be studied with social media analytics tools. Furthermore, a weighting system for this interaction graph is proposed. This system was studied in order to take in account the different nature of several possible interactions. To evaluate our proposals we used a real group of Facebook users. The members of this group were surveyed in other to i) determine the "real" interaction graph; ii) to retrieve useful indications in the weight assignment process. Results show that by looking at interactions instead of friendships it is possible to discover more useful information. The interaction graph gives a more

complete picture of how users communicate to each other. Moreover, our results confirm the fact that active users are key players in a network.

Per aspera ad astra

Contents

List of Figures

List of Tables

3

Introduction

3.1 Motivation

Global criminals are now sophisticated managers of technology using ICT tools in a careful and planned way. Central networked intelligence and coordinated knowledge are fundamental assets available also to non-terrorist criminal organizations. In their investigations, homeland security and intelligence analysis communities typically exploit open source information, which comes from publicly available sources. They rely on valuable instruments to perform intelligence activities and investigations, the so-called Open Source Intelligence (OSINT).

Due to the widespread use of the Internet, one of the most important source of OSINT is the Web and, in particular Social Media (SM). Social media contains an enriched set of data and metadata that can be useful in the intelligence activities. In order to collect, monitor, analyze, summarize, and visualize social media data tools and frameworks have been developed in the so-called field of Social Media Analytics (SMA), a subfield of network sciences that focuses on networks that connect people or social units (i.e., organizations, teams) to one another.

By exploiting technology, frameworks, and tool sets from SMA, Social Media Intelligence (SOCMINT) aims to derive actionable information from social media for applications that can benefit from the “wisdom of crowds” through the Web. The term, coined in a 2012 report [41], refers to *a wide range of applications, techniques and capabilities exploiting social media data in intelligence*.

Although SOCMINT is used not only in the investigation and crime fighting environment, but also for marketing and industrial scopes, nowadays a large amount of data processing is still conducted manually; this results in a waste of both human and time resources. There are software services to perform automatic processing,

but the functionality and degree of automation are still immature and limited, especially considering the vast amount of data today available on social media.

Therefore it is clear that there is a growing need of instruments helping in SOCMINT activities from different perspectives; from the collecting and crawling phases to the analysis and visualization processes. Possible solutions have to take in account several limitations both ethical and technological.

First of all, we must to take in account that types of data manipulated in these activities are strictly dependent on privacy settings; this means that, for instance, an analysis of a friendship graph could be not possible because of the lack of the (whole) dataset.
Furthermore, due the basic idea of the OSINT paradigm, only public information can be used in the process; therefore a dependency of the dataset on privacy settings of the analyzed social mediashould be avoided. By looking at the technological aspects of SOCMINT process, there are great challenges due to the huge quantity of data involved; some of them include data storage and analysis.

3.2 Contribution

The aim of this thesis is to provide a contribute in the field of SOCMINT, especially in the analysis phase. The questions we want to answer are the following:

- How can a Law Enforcement Agency (LEA) successfully exploit social media data?
- Which type of analysis could be useful in an intelligence environment?

A top-down approach is taken to study the SOCMINT: first we study the problem of how to usefully exploit and deal with social media data in intelligence activities and propose an approach to address the issue; then we move on a more specific area: the problem of how to analyse social media data in order to provide to the LEAs useful results and provide, more in depth, an analysis of Facebook data. Third, we improve the analysis proposed by describing a method which takes into account the nature of Facebook data.

Summarizing, our main contributes are the following:

- we design a general framework for the management of social media data;
- we propose an algorithm to extract interactions graphs from Facebook data;
- we propose a weighting system for the interaction graph

that takes into account different types of possible interactions and provides more accurate results.

The proposed framework deals with two important aspects of a social media pipeline. We defined this pipeline as a workflow beginning with the retrieving of raw data from social media sources and ending with analyzed information. In order to easily deal with different types of social media, the framework has been designed with a modular architecture. This enables the extension of the framework itself in a very easy way. The framework is composed of three core modules:

- a module responsible of retrieving raw data from different social media sources;
- a module that performs a first processing of raw data;
- a module that analyzes crawled data and returns, as output, analyzed information.

The first module represents an interface between the framework and the social media sources. It is able to deal with different mechanisms provided by social media platforms in order to access their data. Typically, these mechanisms are implemented using the Application Programming Interface (API) paradigm. These APIs specify a set of functions or routines implementing the interaction between the software requiring data and the platforms providing data. The advantage of using this mechanism is twofold: from a technical point of view, the requesting part does not have to care about the underlying structure of the accessed platform. Regarding legal aspects, APIs pledge the compliance to privacy and technical restrictions implemented in the platform.
The second module, responsible of processing raw data, is designed to deal with multiple formats and types of data that can be retrieved. This phase is fundamental to guarantee a faster and more accurate analysis.
The third is the analysis module, the core of the framework. The general idea is that different analyses can be performed, depending on the type of data and on the goals of the intelligence activities to be conducted. Every type of analysis is performed by a single submodule. In this way, additional analyses can be easily added and executed.

The second main contribute of this thesis to the SOCMINT studies is *an algorithm for the extraction of interaction graph from Facebook data*. The interactions network, in this case, will be a graph built considering only the interactions between users, not only between friends; such interactions are derived from actions performed by these users in public spaces of social media. The main advantage of our proposal is that the algorithm does not require access to private information of the involved users. The main problem occurring with friendship networks, as it typically happens in these kind of analysis, is related to the privacy. Due to the privacy settings managed by users, a lot of interesting information (e.g., the list

of friends) is often not public. To overcome this problem, we chose to use data in spaces in Facebook that are "public-by-design". In this way the two problems given by the privacy settings can be managed: the presence of black holes in the dataset and the compliance to ethical and privacy requirements.

To improve the designed algorithm we also defined, as third contribute, *a weighting system for the interactions graph*. This system was designed considering both theoretical and practical considerations. From a theoretical point of view we considered statistical data to pinpoint most valuable actions. From a practical point of view, we surveyed a group of users about their opinions on "strong" interactions with other users.

3.3 Thesis Overview

The first part of this thesis contains theoretic elements related to the new emergent area of studies in social media: the Social Media Intelligence. Chapter 4 describes the fundamentals of this field of research, by providing elements on how to apply the Social Media-related knowledge to intelligence activities in order to fight organised crime; it also describes the adoption and use of this knowledge by organised crime. To underline the contribute provided by this thesis we also introduce in Chapter 5 a state of art regarding both the advances in the field of SOCMINT in general and in the analysis of Facebook data and, in Chapter 6 a general overview of the architecture.

The second part of this work describes the three above-mentioned contributes, one per chapter. Chapter 7 describes our proposed solution to manage a social media pipeline. An overview of the work is provided in order to better explain the general nature of our studies. A case study, based on Facebook, is then used to illustrate the implementation of the proposed solution. Chapters 8 and 8.3 describe the algorithm for the extraction of the interaction graph from Facebook data and its weighting system. After a first introduction to the concept of interaction graph, we discuss the same concept in the Facebook environment and present all the possible interactions. We then describe the principle of functioning of the algorithm in both general and Facebook cases. To improve the quality of obtained results we propose a weighting system for the interactions graph.
In Chapter 9 we show the obtained results; we choose as proof of concept a Facebook group and we run the algorithm for the interaction graph twice. The first time we study the graph resulted from the non-weighted version whereas the second time we study the same dataset using the weighting system. Chapter 10 discusses the obtained results and draws the conclusions.

4

Background

4.1 Open Source and Social Media Intelligence

Open Source Intelligence can be defined as the retrieval, extraction and analysis of information from publicly available sources, namely Open Source Information (OSINF), opposed to closed or classified sources.

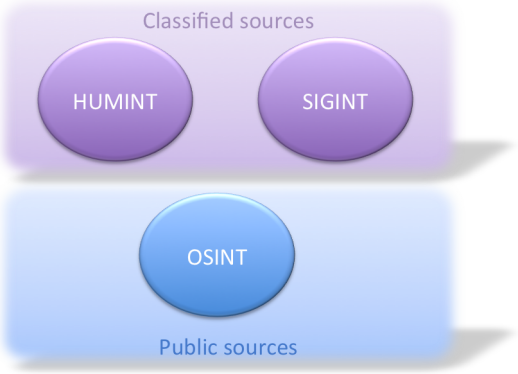


Figure 4.1: Traditional Intelligence Sources

As stated in [3], it is possible to distinguish three intelligence sources (Figure 4.1). HUMINT is intelligence derived from usually clandestine human sources, typically protected with high levels of secrecy. SIGINT is intelligence derived from signal intercepts (e.g. wire taps). OSINT derives from non classified and public sources. Differently from traditional approaches, the sources of OSINT are pub-

licly accessible and have the properties of openness and massiveness. This is, at the same time, a big advantage and a potential disadvantage; the open nature of the data can result indeed in inconsistency and lack of information and validation.

Traditional intelligence management process was most concealed and required massive human effort. It also has the disadvantages of rarity and danger. Therefore OSINT emerged as a major intelligence collection and analysis approach. Large multinational companies, banks and various industries are increasingly relying on business intelligence for decision making and for protecting assets and staff. However it is not just the business world which is moving towards open sources for current awareness and insight. Military organizations, including NATO, recognizes open sources as strategic, cost-effective and rapid intelligence sources [46].

One main advantage of OSINT for Law Enforcement Agencies (LEAs) is that information derived from it can be shared with other agencies. As a result, there is a growing number of OSINT-related services providers in the commercial sector which markets OSINT tools.

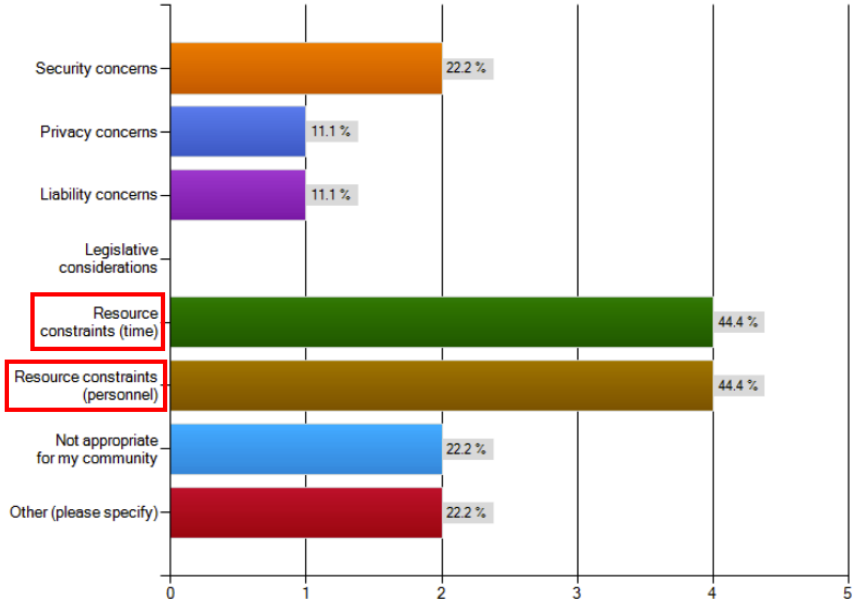


Figure 4.2: Barriers in the use of Social Media by LEAs

Today, the majority OSINT processing is conducted manually: this requires massive human effort and is time consuming. Authorities, such as Law Enforcement Agencies (LEAs), and companies interested in the exploitation of OSINF, have to spend resources and time in order to perform processes such as the collection of data and basic analysis. This is confirmed by a survey conducted by IACP Center for Social Media in 2013[1], on cybercrime and the use of social media by LEAs. Figure 4.2 shows the answers provided by the LEAs to questions about which are the barriers to the use of social media in their environment. Resource constrains, in term of both time and personnel, are the main problems for the 44% of surveyed LEAs.

Automatic processing of OSINT is then unavoidable for modern applications. Although there exists software services to aid such automatic processing, the functionality and degree of automation are still immature and limited [58].

A general process in order to deal with OSINF is shown in Figure 4.3. This can be broken down into five technical processes:

1. Collection: in this phase information retrieval is performed;
2. Processing: information is extracted from gathered data;
3. Analysis: information is analyzed with techniques for, as example, trend or link analysis;
4. Visualization: analyzed data is visualized;
5. Collaboration: knowledge derived from previous phases can be shared.

Nowadays the main source of OSINF, even for intelligence goals, is Internet. In the big sea of Internet a key role is played by social media. Social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that enables the creation and exchange of User Generated Content (UGC) [29].

Web 2.0 (the term was firstly used in 2004) represents the ideological and technological base of social media. It is a new way in which software developers and end-users started to utilize the World Wide Web; content and applications are modified by all users in a participatory and collaborative fashion. User Generated Content (UGC) are the several forms of media content, publicly available, created by end-users. The Organisation for Economic Cooperation and Development defines a UGC as a content fitting following requirements: i) it is published either on a publicly accessible website or on a social networking site accessible to a selected group of people; ii) it shows a certain amount of creative effort; and iii) it has been created outside of professional routines and practices [49].

[1] http://www.iacpsocialmedia.org/Portals/1/documents/2013SurveyResults.pdf

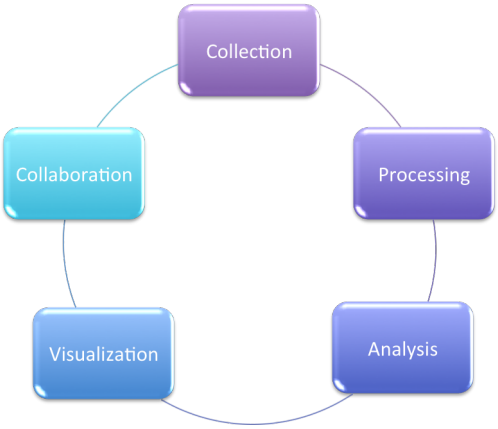


Figure 4.3: How to process OSINF

For many individuals, social media are the unique information source to deal with information, to network and share their opinions and their expertise in several kinds of dynamic dialogs [62].

For homeland security and intelligence analysis communities, social media represent an opportunity to study, for example, terrorist group behaviour, including their recruiting and public relation schemes and the grounding social and cultural contexts.

Social media contains an enriched set of data and metadata, that can be treated systematically in data- and text-mining literature [2]. To collect, monitor, analyze, summarize, and visualize social media data, tools and frameworks, usually driven by specific requirements from a target application, have been developed in the so-called field of Social Media Analytics (SMA).

By exploiting technology, frameworks, and tool sets from SMA, Social Media Intelligence (SOCMINT) aims to derive actionable information from SM data for applications that can benefit from the “wisdom of crowds” through the Web. The term, coined in a 2012 report [41], refers to a wide range of applications, techniques and capabilities exploiting social media data. The relation between SOCIMINT and OSINT is not trivial. Even if some analysts describe SOCMINT as a branch of OSINT, some others underline that SOCMINT is defined, not by the openness of the information on which it is based, but by its existence on a social media platform.

Knowledge about criminal networks has important implications for crime investigation and the anti-terrorism campaigns. However, the lack of advanced, au-

tomated techniques has limited law enforcement and intelligence agencies ability to combat crime by discovering structural patterns in criminal networks.

4.2 Organised Crime and Social Media

Organized crime involves multiple actors. For quite some time, organized crime groups were considered as groups of at least three offenders, involved in different types of offences with the primary objective to gain financial profits [48]. To define organized crime [17] proposed a number of common characteristics:

- collaboration of more than two people;
- acts of serious criminal offenses (suspected);
- determined by the pursuit of profit and/or power;
- each having their own appointed tasks;
- for a prolonged or indefinite period of time;
- using some form of discipline and control;
- operating across borders;
- using violence or other means suitable for intimidation;
- ising commercial or businesslike structures;
- engaged in money laundering;
- exerting influence on politics, the media, public administration, judicial authorities, or economy.

As suggested in [11, 10] it is possible to identify three categories of organized groups that exploit advances in information and communications technologies (ICT) to infringe legal and regulatory controls: i) traditional organized criminal groups which make use of ICT to enhance their terrestrial criminal activities; ii) organized cybercriminal groups which operate exclusively online; and iii) organized groups of ideologically and politically motivated individuals who make use of ICT to facilitate their criminal conduct.

As an example, the number of terrorist-linked websites has grown from about 15 in 1998 to more than 4500 today ([11]). These sites use slick multimedia to distill propaganda whose main purpose is to i) enthuse and stir up rebellion in embedded communities, ii) instill fear in the enemy and fight psychological warfare. Anonymous communication between terrorist cells via bulletin boards, chat rooms and email is also prevalent.

Global criminals are now sophisticated managers of technology. The sophistication and high level of knowledge shown by cyberterrorists and criminal networks require more efforts to be put in place by governments, companies and citizens.

Terrorist threads or malware expansion have in common the careful and planned using of ICT tools.

"Modern organized crime has abandoned the top heavy structure of dons, capos, and lieutenants made famous in The Godfather. Most of today's gangs, along with Al Qaeda and other terrorist groups, are loosely affiliated cooperative networks and are as likely to recruit website designers and hackers as they are thugs and enforcers." [25].

It is important to take in account, as remarked in [10], that traditional organized crime groups should not be confused with organized cybercrime groups that operate exclusively online. Objectives targeted by organized crime include extortion, industrial espionage and monitoring the traffic flow in order to be alert against defensive counterattacks [16]. As example, the Zeus Trojan gang was eventually arrested in September 2010, after stealing several millions from USA and UK bank accounts. Sophisticated Trojan worms as Zeus may change over fifty times in a single day, depending on the conditions in which they are hosted, and sending automatically information on the general security conditions of the host to remote criminal servers. Calculate risk levels and keeping tuned the programming according to them is not only a common practice among the police and security companies. It is also a common practice among cybergangs.
Twelve years ago, industrial piracy and traditional crime organizations were taking advantage of the Internet in a more limited scale, leaning on the black market and on a high hierarchical and tight structure, daily controlled. These types of criminal behavior did not disappear and keep going on.

Today Internet is not only a tool, but the condition and natural environment of the organized crime. Extremist and terrorist groups use the Internet for a myriad of purposes, including the dissemination of propaganda, the recruitment of new members and the development of operational planning. Online activity is a critical part of almost every national security investigation. By 1999 nearly all known terrorist groups had established a presence on the Internet.

According to the advance of the Web 2.0, online child pornography [8], prostitution [14], terrorism and all sorts of extortion and aggressive behavior have been fueled by the expansion of the social web. Popular social networking websites are another means of attracting potential members and followers. These types of virtual communities are growing increasingly popular all over the world, especially among younger demographics. Youths are especially targeted for propaganda, incitement and recruitment purposes. Predominately-Western online communities like Facebook, MySpace and Second Life and their Arabic equivalents are being used more and more by terrorist groups and their sympathizers. Counter-terrorism

expert Anthony Bergin says that terrorists use these youth-dominated websites as recruitment tools, "in the same way a pedophile might look at those sites to potentially groom would-be victims" [53].

Unfortunately, there are no detailed empirical research into how extremist and terrorist groups have reacted to the rise of social media. Anyway, organised crime shifted from text-heavy traditional websites to social networks built around interactive forums allowing the sharing of mixed media (often where leaders posted stories and steered discussions) in the mid 2000s. Recent analysis suggests that since late 2000s the activity has increasingly shifted to social media platforms. For instance, according to Aaron Zelin ([61]), "it is only a matter of time before terrorists use Twitter and Instagram as part of ongoing operations". On his analysis of Jihad, Zelin charts an increase in activists using Twitter as a tool of communication, motivated perhaps by the need to appeal to a younger demographic that prefers this medium.

Figure 4.4 shows the different four phases in which Jihad's media have been disseminated since 1984 (phase 1 is not in the Figure because only newspapers and similar were available. The dates roughly correspond to the adoption of a new medium for distributing information.

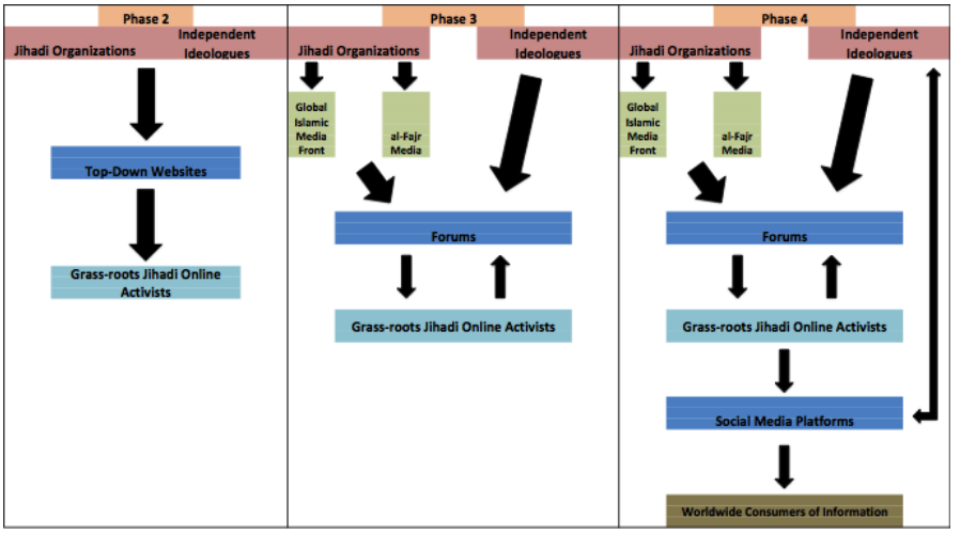


Figure 4.4: The evolution of Jihad media adoption
Phase 2: mid 1990s - Phase 3: mid 2000s - Phase 4: late 2000s

Moreover, a recent report[2] from the MEMRI (Middle East Media Research Institute) documented the use of Instagram by al-Qaeda leaders to share images and quotes, glorify imprisoned fighters, and disseminate images of dead "martyr".

According to [39] social media is also used by neo-Nazi groups to redirect users to content hosted on external websites. Indeed, it is this ability to share news items, original articles and essays and tribute videos that is perhaps key.

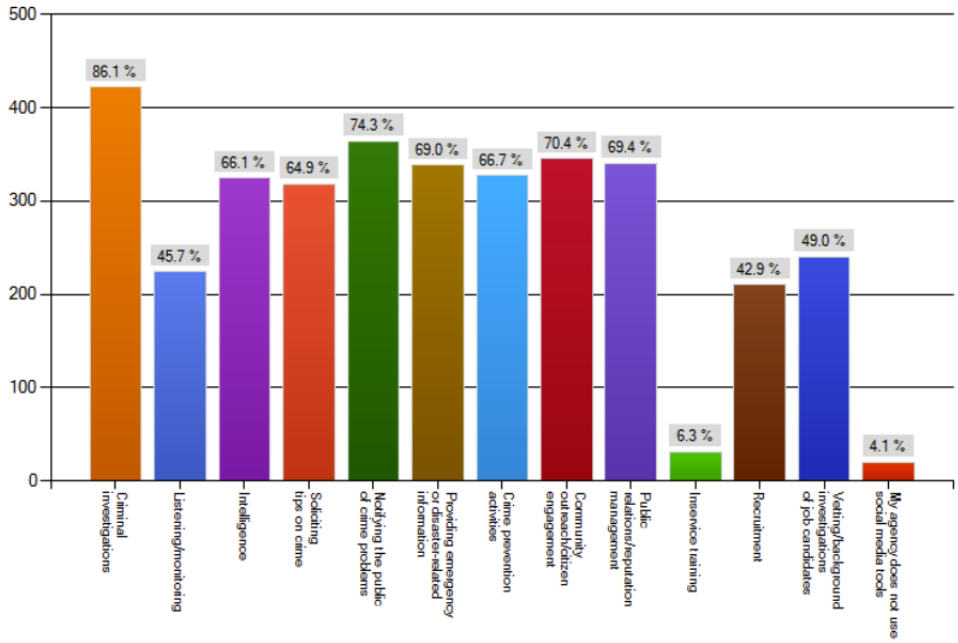


Figure 4.5: Scopes of the use of Social Media by LEAs

It is possible then to state that social media use is affecting law enforcement activities: as emerge from the report conducted by IACP Centre for Social Media, nowadays the 95.9% of agencies surveyed use social media and, among those that already use these instruments in their activities, the 86,1% use it for criminal investigations (see Figure 4.5).

Social media are also everyday more relevant in public disorder policing. In both the August 2011 riots in the UK, and in Vancouver following the Stanley Cup the same year, a common tendency was identified. During the early stages of

[2] http://www.memri.org/martyrs-instagram.html

disorder, participants and uninvolved observers recorded and shared information about the event. As the disorder increased, information describing the apparent impunity of the rioters, visibly shared on social media, may have escalated the disorder further.

4.2.1 Privacy concerns and limitations

The use of social media to collect information by LEAs presents challenges to the existing legal and ethical frameworks that manage the various types of harm that can result from intelligence gathering. Like all intelligence work, SOCMINT must be carried out according to legal frameworks. The legislation that covers the collection and use of private information differs from one country to another.

Furthermore, studies suggest that users have increasingly become aware of the privacy risks and reacted using more restrictive privacy settings for their UGC. A recent study, described in [18], of 1.4 million Facebook users in New York has shown that in 15 months between 2010 and 2011 users who kept key attributes on their profiles private rose from 12 per cent to 33 per cent (see Figure 4.6). In the same way the nature of the terror threat is likely to evolve in future, and could include groups that are expert in various counter-surveillance and "sous-veillance" techniques (monitoring agents). An example of this trend is the app for smartphone Vibe. This platform become popular during the Occupy Protest movement in 2011, because allows users to send short anonymous messages to users within a pre-defined geographical proximity; messages are automatically deleted after a pre-determined period of time.

From the opposite point of view, companies providing social media services, especially those social networks providers like Facebook, exploit the huge amount of information they own about users and their activity to build behavioral and targeted advertising. Then, due to these business driven reasons, social media platform providers are reluctant to share information about users and the usage of the network [26, 33].

For this, social networks are implemented as a black box. Facebook, for example, implements several rules, some behavioral (e.g. terms of use prohibits data mining, regardless of the usage of data extracted) and some technical mechanisms (e.g the friend list is dispatched through a script that interact asynchronously with the Web page, preventing naive techniques of data extraction) to avoid an excessive flow of information.

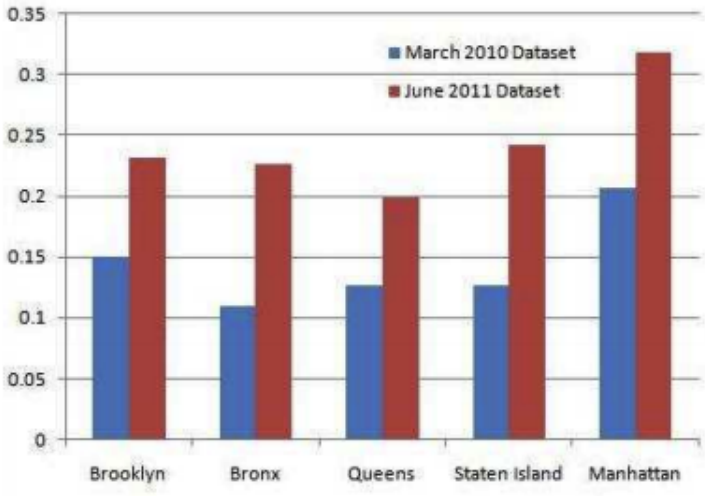


Figure 4.6: Numbers of Facebook hidden friends in NYC

4.3 CAPER project

We started developing our system for the CAPER FP7 European project. The main objective of the project is to build a common collaborative and information sharing platform for the detection and prevention of organised crime in which the Internet is used (e.g. sale of counterfeit or stolen goods, cybercrime) and which exploits Open Source Intelligence.

The end users of the platform are the European Law Enforcement Agencies (LEAs). Information that can be retrieved freely online are used as input of different types of analysis; textual analysis for the individuation of named entities in texts, biometric analysis for the detection of people in video/images, audio analysis for the recognition of speech. etc.

One of the possible source of information, especially related to the analysis of interactions between people, is represented by Social Media (SM), in particular by Social Networks (SN). Using public user generated content (UGC) it is possible to study the the interactions between users in terms of frequency, strength, duration, etc.. The results of the analysis, intersected with the results of other analysis, can provide to the LEAs a better understanding of the situation under investigation, highlighting relations, topic of discussions, people involved.

5

State of the Art

A variety of tools is regularly used in both the tactical and strategic analysis process of intelligence. Examples are case comparison analysis, profiling of the offense and of involved people (the victim or the offender), spatial analysis (GIS), crime pattern analysis, event flow charting (i.e. timeline of relevant events), template matching, and telephone toll analysis [34].

The standard procedure for investigating collaborations of criminal activity is focused on the profiling of relations between criminal actors, typically performed using the link analysis [28]. This method reviews criminal actors, events and activities and visualizes the links between entities in a graph. The goal of these kind of analysis is the discovery and identification of patterns of activity, criminal roles, and key players. In literature there are examples of link analysis approaches to the study of crime networks. In [56], for example, the authors proposed a link analysis technique that uses shortest-path algorithms, priority-first-search (PFS) and two-tree PFS, to identify the strongest association paths between entities in a criminal network. The evaluation was conducted comparing the PFS algorithms with crime investigators' typical association-search approach.

It is important to underline that i) criminal investigators mainly use the visual mapping tools without bothering much about the mathematical considerations and social mechanisms underlying such networks ([30]), and ii) the link analysis method disregard that perceptions of relational patterns are affected by the layout of graphs. As demonstrated in [35, 36] people tend to consider actors in the centre or at the top of a graph as most important.

According to [13], the failure to analyze networks in a more objective manner is one of the major reasons why law enforcement and intelligence can fail. The Social Network Analysis (SNA), as theoretical and methodological paradigm for

sophisticated examining of complex social structures, is considered as the scientific equivalent of link analysis. Then SNA can help to systematically uncover clandestine and adversary networks.

Sophisticated network analysis methods need to enable investigators to identify positions of power and to attribute these to specific individual traits or to structural roles that these individuals fulfill. Social network introduce dynamics into the rigid and 'frozen' understanding of social structures that traditional organisational diagrams convey. Exploiting these techniques, processes of recruitment become clearer by looking at previous connections, as does the transfer of knowledge and criminal innovations. Innovating criminal analysis alone however will not suffice. Ultimately, controlling organized crime can only be done successfully through more flexible modes of organisation and operation, thus creating effective law enforcement and intelligence networks to deal with criminal networks [30].

An overall approach to the use of SNA in intelligence is described in [57]. The authors implement a framework including four stages: network creation, network partition, structural analysis, and network visualization. The system, called CrimeNet Explorer, incorporates several techniques, such as hierarchical clustering, social network analysis methods, and multidimensional scaling.

Another example of the exploitation of SNA in intelligence activities is presented in [45]. The authors propose a model for targeting criminal networks based on Borgatti's key player assumption [6, 5]. This approach attempts to identify central actors of a networks measuring explicitly the contribution of a set of actors to the cohesion of a network. The added value of the approach in [45] is the incorporation of the weights of network actors and their associations. A particular system for attribute or link weights is used; the methodology is compatible with any numerical system used by LEAs.

A recent work [20] explains the application of social network theory in Dutch law enforcement. The authors present a case study of the 'Blackbird', a crime network involved in the wholesale cultivation of cannabis. In this case SNA is combined with crime script analysis: using a mix of quantitative and qualitative analysis, the topology of the 86-strong Blackbird network is laid out and its substructures and key individuals exposed.

In [55] the authors exploit the concept space approach, clustering technology, social network analysis measures and approaches, and multidimensional scaling methods for automatic extraction, analysis, and visualization of criminal networks and their structural patterns. In pairs with Tucson Police Department authors conducted a case study on gang and narcotics criminal enterprises. The proposed approaches could detect subgroups, central members, and between-group inter-

action patterns correctly most of the time. Moreover, the system extract the overall structure for a network.

Social networks, in the large sense of the term, have been a subject of intensive study since the 1930s. The structure of social networks is important then not only from the perspective of the individual, but also from that of the society as a whole. However, uncovering the structure of social networks has been constrained by the practical difficulty of mapping out interactions among a large number of individuals [42].

Social scientists have ordinarily based their studies on questionnaire data [51]. In the late the 1990s, physicists became interested in large-scale social networks [52, 1] revealing how individual microscopic interactions translate into macroscopic social systems. Based on these assumptions, studies concerning interactions on social media has emerged. In literature it is possible to find several examples of studies related to the role of the activity or communication networks. In [32] an undirected communication network is created and analyzed from data collected from MSN instant messenger services.

The first attempt to study the role of interactions on social network in comparison of the friend relationship network is presented in [12]. The authors investigate the case study of Cyworld and construct a network from comments written in guestbooks in which a node represents a user and a directed edge a comments from a user to another. We have analyzed structural characteristics of the activity network and compared them with the friends network. The activity network has shown topological characteristics similar to the friends network, but thanks to its directed and weighted nature, it has allowed us more in-depth analysis of user interaction.

In literature is possible to find a class of socially enhanced applications that leverage relationships from social networks to improve security and performance of network applications, including, for example, spam email mitigation [22], Internet search [38], and defense against Sybil attacks [60]. In all the above-mentioned cases, interactions between friends are critical to improving trust and reliability in the system. Even if these studies exploits social network connectivity statistics, social psychologists have long observed the prevalence of low-interaction social relationships similar to Milgram's "Familiar Stranger" [37]. As support of the hypothesis that social links often connect acquaintances with no level of mutual trust or shared interests, researchers described in [19] that users of social networks often use public display of connections to represent status and identity.

In the social media business environment measuring the Facebook interactions is considered a must for successful social media marketing. Facebook is a bidirec-

tional medium, just like virtually all social media channels. Therefore every kind of feedback, meaning fans communicating by likes, shares and comments directly on page owners' wall posts, needs to be monitored. Furthermore, if activated, fans can also directly post to the page's wall. Consequently, interaction naturally turns into an important indicator of success.

Another example of studies related to interactions on social media is the work presented in [21]. This study aims to investigate the relationship between use of Facebook and the formation and maintenance of social capital. The authors surveyed a group of high school students and collected information two types of information: demographic and other descriptive variables (e.g. gender, age, year in school, local vs. home residence, ethnicity) and Facebook usage measures, such as time spent on Facebook and items designed to assess whether Facebook was used to meet new people or to establish an online connection to pre-existing connections. Their results demonstrate a robust connection between Facebook usage and indicators of social capital. The strong linkage between Facebook use and high school connections suggests how social networks help maintain relations as people move from one offline community to another one.

The main problem of these approaches is the basic assumption that all online social links denote a uniform level of real-world interpersonal association. While most social networking sites allow only a binary state of friendship, not all links are created equal. A recent study, presented in [23] has demonstrated that, even if social media treats all users, real friend or stranger, in the same way, the "strength of ties" between users varies widely. In order to distinguish between these strong and weak links, researchers have suggested examining, instead of the friendship network, the activity network. The activity network is the network formed by users who actually interact each others using one or many of the methods provided by the social networking site. Initial studies on activity networks, such as presented in [54, 12] demonstrate that an activity network is structurally different from the social network. Moreover, as underlined in [50] the level of interaction between two individuals can vary over time and it is possible to examine how the varying patterns of interaction affect the overall structure of the activity network. The authors of the paper studied the evolution of activity between users at both, the microscopic level (investigating how pairs of users in a social network interact), and the macroscopic level (examining how the varying patterns of interaction affect the overall structure of the activity network).

Weak points on the state of the art can be summarized as follow:

- Social media intelligence is a recent field of research and, today, tools and techniques for the exploitation of social media in intelligence activities are few.

Case studies already present in literature involve directly LEAs and are conducted on their closed data.

- Dealing with social media poses challenges in terms of privacy and storage. Application not requiring access to users private data and that guarantee compliance with privacy and ethical requirements are needed.
- Typically the studies of interactions and related activity network are targeted to the comparison with the friendship network.

6

Overview of the architecture

The final goal of our work is to provide instruments useful to LEAs in intelligence activities based on SM data. The contributions of our work in the Social Media Analytics for Intelligence field covers different aspects. First of all, we designed a framework to manage what we can define a general *social media pipeline*. A social media pipeline is a workflow starting with the retrieving of raw data from SM sources and ending with analyzed information.

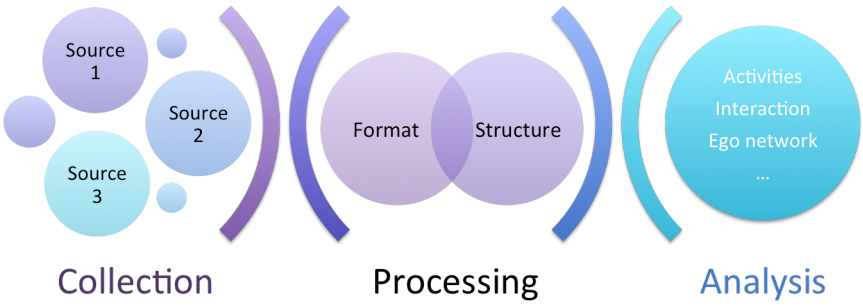


Figure 6.1: Social Media pipeline

As reported in Figure 7.1 this pipeline is composed of several steps. The first phase, collection, is the retrieving of raw data from social media sources. This phase, as explained in detail in the next section, is not trivial, even if the mechanisms provided by the platforms can be exploited. The resulted data set is influenced by different factors, such as the choice of seeds, node selection algorithms, and the sample size. These factors may introduce biases and further contaminate or even skew the results [59]. Given the huge dimension of graphs typically involved in these kind of analysis, during the crawling phase it is important to take care about some aspects to be addressed: i) which sampling method to use, ii) how small can the sample size be, and iii) how to scale up the measurements of the sample (e.g., the diameter), to get estimates for the large graph [31]. An in depth discussion concerning the crawling of SM is out of scope of this thesis.

After the collection, data retrieved by different SM platforms need to be stored and processed in order to be analyzed correctly. First of all a pre-processing and cleaning phase is mandatory. These steps are addressed to solve problems related to both format and storage issues. Regarding the format issue, SM platforms typically provide data in JSON format but, due to the different nature of SM involved, data retrieved varies in different fields. To analyze these data it is then mandatory to try to align JSON fields as much as possible. Moreover, the storage of huge amount of data, especially in a real-time analysis environment, is not a trivial task.

The third phase of a general SM pipeline is the analysis. During this phase, different types of analysis can be performed, typically depending on the goals. Analysis that can be carried out are, for example, the analysis of the friendship network or the analysis of interactions between users. It is at this step that we provided another contribute: we designed an algorithm for the extraction of interactions networks from SM data, firstly introduced in [54].

We chose to focus our attention on interactions, instead of friendship networks, for several reasons. The main problem in dealing with friendship networks, as typically happens in these kind of analysis, is related to the privacy. Due to the privacy settings managed by users, a lot of interesting information (e.g. including the list of friends) are often not public.
To overcome this problem, we chose to use data in spaces in SM that are "public-by-design"; pages or groups on Facebook, comments on public posts on blogs, etc. In this way two problems deriving from privacy settings can be managed; the presence of black holes in the dataset and the compliance to ethical and privacy requirements. Moreover, even if the friendship networks is a good indicator of different social aspects, people interacting each other on some topic or discussion,

including those related to illegal activities, are often not "friends". Consequently their relations do not emerge, for example, from analysis of the ego networks.

To improve the designed algorithm we also defined a weighting system for the interactions network. This system was designed considering both theoretical and practical considerations. From a theoretical point of view, we considered statistical data to pinpoint most valuable actions. From a practical point of view we surveyed a group of users about what they consider as most "strong" actions.

All the contributions have been designed to be as much general as possible. This means that regarding, for example, the framework for the management of a SM pipeline, all modules are general, independent from the specific SM used. The same is valid for the interactions network and the related weighting system.

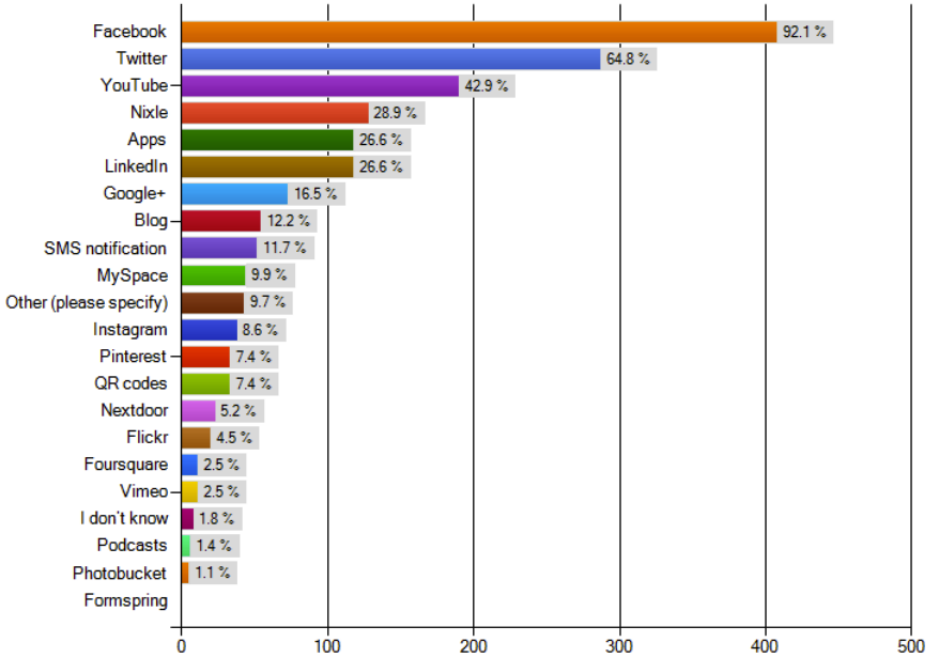


Figure 6.2: Social Media currently used by LEAs

In order to provide a more clear explanation of our contributes and show significant results, we choose, as central idea, a case study based on Facebook. Among the available SM we chose Facebook for two main reasons: the first one is related to the target of our studies, the LEAs. According to a survey conducted in 2013

by IACP Social Media Center [1] and as shown in Figure 6.2, Facebook is the most used social media among LEAs that already use social media for their activities, representing the 95.9% of agencies surveyed.

Even looking at LEAs that do not use yet social media in their activities, Facebook is the first preference, as shown in Figure 6.3.

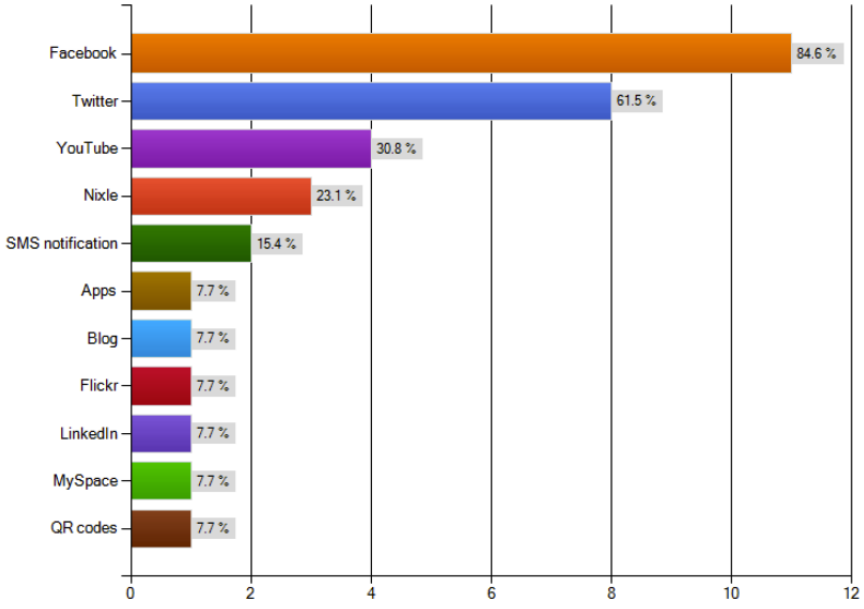


Figure 6.3: Social Media considered for the adoption by LEAs

Furthermore we chose Facebook as proof-of-concept due to his widespread adoption that corresponds to a huge amount of open data available. As shown in Figure 6.4, today Facebook is the most used social media in the world. According to the statistics of the 1st January 2014, the worldwide number of Facebook active users was 1,310,000,000[2], with an increment, comparing to the previous year, of the 22%.

User Generated Content(UGC) can be shared on Facebook not only using user's profiles (that are more often private), but also using public spaces provided

[1] http://www.iacpsocialmedia.org/Portals/1/documents/2013SurveyResults.pdf
[2] http://www.statisticbrain.com/facebook-statistics/

by the platform: pages, groups and events. In these public areas, users communicate each other using several types of interactions: posts, comments, likes, shares. As remarked by the statistics, these instruments are widely used and the total number of Facebook pages, at the time of this writing, is 54,200,000.

We will explain our contributes in each Chapter by providing a general discussion first, followed by a detailed description of the application to the specific case study.

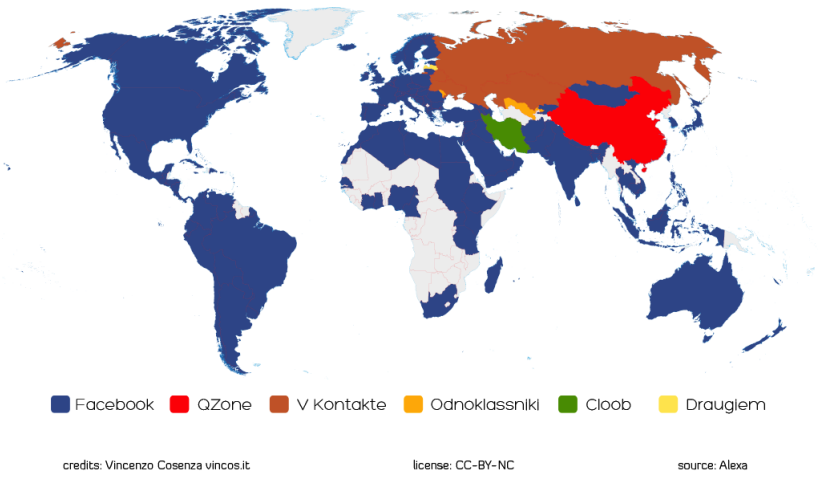


Figure 6.4: World use of Social Networks

7

Managing a Social Media pipeline

In this chapter we describe our first contribute: a framework for managing a generic social media pipeline. A social media pipeline is a workflow starting with the raw data retrieval from social media sources and ending with the analyzed information.

Due to the fact that we focus on the exploitation of social media analytics on intelligence activities, a parallel comparison with a general OSINT process, as stated in [3], can be helpful.

As shown in Figure 4.3, the proposed framework covers the first three phases: collection, processing and analysis of data. The framework is modular, therefore visualization modules can be added easily.

As reported in Figure 7.1 this pipeline is composed of several steps.
The first phase, collection, is the retrieving of raw data from social media sources. After the collection, data retrieved by different SM platforms need to be stored and processed in order to be analyzed correctly. First of all a pre-processing and cleaning phase is mandatory. These steps are addressed to solve problems related to both format and storage issues. Regarding the format issue, SM platforms typically provide data in JSON format but, due to the different nature of SM involved, data retrieved varies in different fields. To analyze these data it is then mandatory to try to align JSON fields as much as possible. Moreover, the storage of huge amount of data, especially in a real-time analysis environment, is not a trivial task.

The third phase of a general SM pipeline is the analysis. During this phase, different types of analysis can be performed, typically depending on the goals. Analysis that can be carried out are, for example, the analysis of the friendship network or the analysis of interactions between users.

Our framework, presented in Figure 7.2 is composed of two distinct modules: the first one, namely the Social Media Capture (SMC), is responsible of the collec-

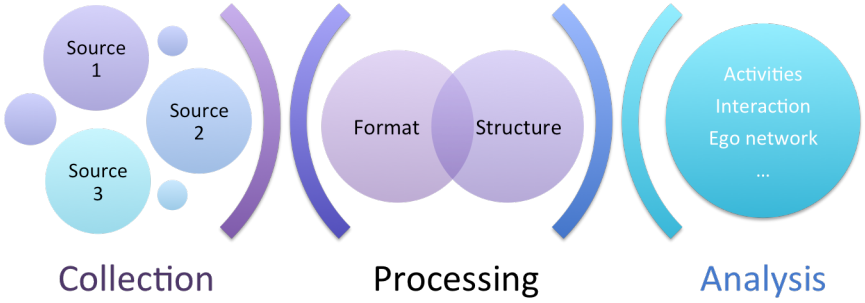


Figure 7.1: Social Media pipeline

tion (phase 1) of information in SM and of the processing (phase 2) of raw data. The second module, called Social Media Analyzer (SMA), is composed of different submodules, performing the analysis (phase 3) of the general process.

The task of retrieving data from social media sources is not trivial due to legal, ethical and technical issues. The resulted data set is influenced by different factors, such as the choice of seeds, node selection algorithms, and the sample size. These factors may introduce biases and further contaminate or even skew the results [59].

First of all, social data is among the most valuable assets to the social media providers, therefore it is hard to get such data. Typically, SM providers implement technical restrictions to control the information flow. Examples of these restrictions are IP-based access, rate limit, and banning politics.

Second, the amount of data that can be crawled includes millions of different entities (e.g. comments, contact lists, profiles, pictures, videos, etc.) in several different formats. Given the huge dimension of graph typically involved in these kind of analysis, during the crawling phase it is important to take care about some aspects to be addressed: i) which sampling method to use, ii) how small can the sample size be, and iii) how to scale up the measurements of the sample (e.g., the diameter), to get estimates for the large graph [31].

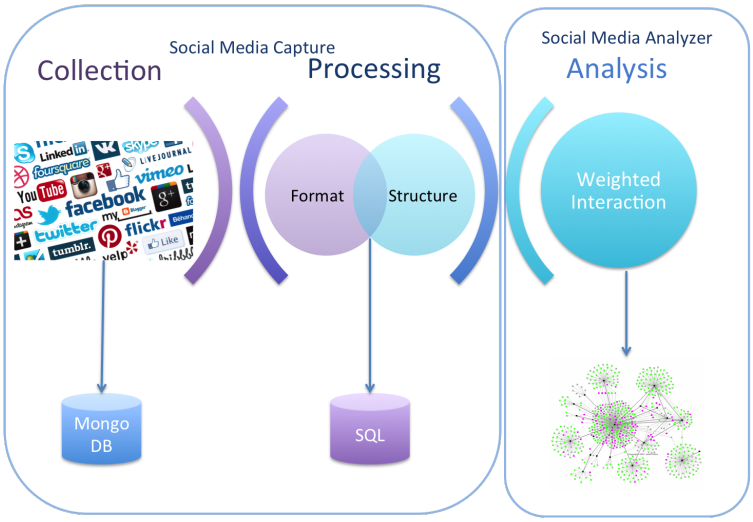


Figure 7.2: Social Media Pipeline Manager framework

In addition to this, to social media users usually is granted the flexibility to customize the layout of their pages, which further complicates the design and implementation of a parser [59]. For these reasons a proper storage support must be used. An example of this problem is reported in [24]. Authors measured the crawling overhead in order to collect the whole Facebook graph in 44 Terabytes of data to be downloaded and handled. This problem is approached by several works in literature, with a special focus on OSNs crawling. Methods to check the goodness of sampling, and scaling laws that describe relations between the properties of the original and the sample are also treated [31]. Moreover, efficiency problem, related to how fast different crawlers discover nodes/links, sensitivity, related to how different OSNs and the number of protected users affect crawlers and bias are also discussed [59]. An in depth discussion concerning the crawling of SM is out of scope of this thesis.

Third, as more users become concerned about their privacy, especially in online social networks, many of them choose to keep private their information; hence they acts as “black holes” for crawlers.

The module we designed for the collection phase of the social media pipeline is composed of three distinct parts:

- An input module for used to determine the type of entities involved in the crawling
- A crawling server that actually performs the retrieving of data
- An output module that stores data

Even if this conceptual model is general and can be applied to every social media, it is important to underline that, especially regarding the crawling server, there is a strong dependence on the chosen source. First of all, it is necessary to verify the presence of retrieving mechanisms provided by the platform.

The input is then necessary to prepare the crawling phase according to the entities involved. With the terms entity we refers to object, typical in social media data, representing different mechanisms provided to the users by the platforms itself and allowing the communication and the interactions between users. Examples of entities can be Facebook pages, groups and events.

Nowadays, a common way to get data from social media are the Application Programmable Interfaces (APIs) provided by the platform itself. APIs are used to specify a set of functions or routines performing the interaction between software component. In this way, there is no need, from the requesting part, to care about the underlying structure of the module providing the service. The advantage of using APIs provided is also related to legal aspects; APIs exposed by social media platforms are compliant to privacy and technical restrictions.

The processing phase is designed in order to guarantee a more general approach. In this phase of the process crawled data are prepared in order to be analyzed. In particular, a mechanism that splits crawled data in different types depending on their nature, was designed. Data can be retrieved from a social media are not only texts, but also multimedia contents (i.e. video and images) or metadata related to the structure of the social media itself. In this phase retrieved content is then separated. In this way a focused analysis, on different types of data, not only aggregate ones, can be performed in parallel using different analysis modules.

The Social Media Analyzer is the module responsible of the analysis on the crawled data. It exploits submodules for each different analysis task and type of data. In this way it is possible to introduce a new submodule with a new analysis and/or data and it is possibile to perform more analysis tasks in parallel. The Social Media Analyzer module is the core of the social media pipeline. It is responsible of the different type of analysis that can be performed. Single analysis can be performed in parallel by single submodules. Analysis can involve different aspects: text extracted from data can be analyzed with text-mining techniques, images can be processed according to image analysis algorithms, links between people can be studied using social networks analysis methods.

7.1 Implementation of SM pipeline - the Facebook Case Study

In our implementation, we exploited the low-level HTTP-based APIs provided by Facebook: the Graph APIs. We manage the retrieving of publicly available data in pages, groups and events. A FB page is a virtual place used by companies, organizations, famous people, to promote their activities and to interact with users/fans. Moreover, these pages represent often a kind of virtual plaza where users exchange their opinion and feeling about a specific topic. A Facebook group is similar to a page, but while pages allow real organizations, businesses, celebrities and brands to communicate broadly with people who like them, groups are closed space for small groups of people to communicate about shared interests.

Table 7.1 details the difference between pages and groups. Events is a feature that lets users organize gatherings, respond to invites and keep up with what their friends are doing. From a technical point of view, also events are public space in which UGC can be shared. Our module is able to retrieve the feed, namely the stream of posts and related comments, published on public space (i.e pages, groups, events) and related general information. We will refer to these three types of public spaces as entities.

	Privacy	**Audience**	**Communication**	**Administration**
Pages	Everyone	Anyone on FB	Page admin	Representative
Groups	Different privacy settings	Groups members	Groups members	Anyone on FB

Table 7.1: Differences between pages and groups on Facebook

Page information and posts are public and generally available to everyone on Facebook. Anyone can like a Page and get news feed updates. There is no limit to how many people can like a page. When a page administrator share posts these appear in the news feed of people who like the page. Pages may only be created and managed by official representatives.

Compared to pages, groups have more privacy settings available. In secret and closed groups, posts are only visible to group members. Group members must be approved or added by other members. Features available are limited according to the group size; then, if a group reaches a certain size, some features can be limited.

By default members receive notifications when someone of the member posts in the group. Group members can participate in chats, upload photos to shared

albums, collaborate on group docs and invite members who are friends to group events. Groups can be created by anyone.

The communication between our module and the Facebook platform is performed via an ad-hoc Facebook application, using credentials of a generic unauthenticated user. A key point on this phase is the storage of the data; due the huge amount of unstructured data that can be retrieved we chose as storage support MongoDB[1] . MongoDB (from humongous, extremely large) is a cross-platform document-oriented database system (NoSQL) that use JSON-like documents with dynamic schema in a specific JSON-like format, the BSON. The advantage of using these technologies relies on a easier and faster integration of data.

The first step to accomplish, in order to perform the crawling data, is the registration of the entity to crawl. At this step several parameters, including the type of the entity and the time interval, are configured. The most important ones are the keywords that will be used as searching parameter in the publicly available data on Facebook. The choice of these keywords is fundamental: this input determines the "topic" of the crawler and therefore of the analysis. Public spaces that can be retrieved in this way are those containing these keywords on their title.
This strategy was chosen because this allow us to crawl data that are more "focused" on the topic. Moreover, a technical reasons have to be considered. Searching also content can result in data belonging to private spaces.

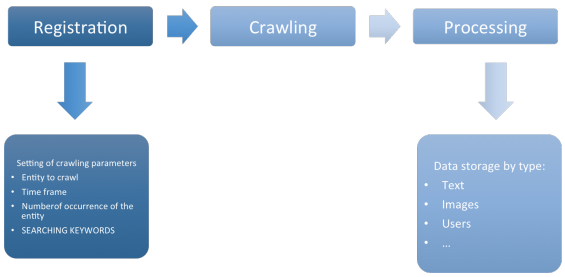


Figure 7.3: Social Media Capture

After the registration, the crawling phase starts. Due to the fact that a crawling operation can be time consuming, a mechanism, such as described in [9], to make parallel crawling is implemented. The crawling is performed using a Snow-

[1] https://www.mongodb.org/

ball Sampling-like algorithm [47]. A snowball sampling algorithm, starting from a central node, collects all possible related data, before passing to the subsequent node. The original version of the algorithm is focused on the crawling of friend's network. In our case, central nodes are posts on the entities or friends and likes, comments and mentions on the post itself.

The output module is in charge of performing processing tasks. These tasks are fundamental in order to enable a correct and as faster as possible analysis of the data. As shown in Figure 7.3, texts and images retrieved are separated from other information. In this way, separated analysis on text and images can be conducted.

In our proposal we present a type of analysis related to interactions between users; a detailed description on how this analysis works is given in the next chapter.

8

From data to the interactions graph

Sociologists define social interactions as particular forms of externalities, in which the actions of a reference group affect an individual's preferences. The reference group depends on the context and is typically an individual's family, neighbor, friend or peer. Social interactions are sometimes referred as non-market interactions to underline the fact that these interactions are not regulated by the price mechanism [44].

The concept of interaction graph was firstly introduced in [54], as a subset of the social graph parametrized by two constants n and t, where n defines a minimum number of interaction events, and t represents a time window during which interactions must have occurred. n and t delineate an interaction rate threshold. Starting from these assumptions, an interaction graph is the subset of the social graph where, for each link, interactivity between the link's endpoints is greater than the rate given by n and t.

Differently from this original definition, we will refer to an interaction network as a graph built considering only the interactions between users, not only between friends, derived from actions performed by these users in public spaces of social media. One of the most valuable advantages of this proposed definition of interaction is the fact that there is no dependency from the friendship graph. In this way we can overcome limitations derived from privacy and ethical reasons; these limitations, such as the need for accessing the profile of users in order to collect information about her friendship graph, influence both, the crawling and the analysis process.

From now on, we refer to an interaction network as a graph, built only with public information and composed of nodes representing Facebook users of pages, groups or event and edges representing possible interactions.

The main difference between the above mentioned study and our contribute rely on the fact that, in the cited research, the authors implicitly state that the majority of user interaction events occur across social links. Due to the constraints posed by our SOCMINT scenario, we skip this assumption because we look at interactions on public places on Facebook.

Considering the crawling process, as remarked also in Chapter 7, privacy settings of the users can transform their profiles as black holes for a crawler module. Concerning the analysis process, the presence of black holes in the dataset to be analyzed could affect the completeness of the obtained results. Furthermore, as stated for example in [54], the simple fact that two users have a friendship relation online is not a confirmation of their real interaction.

8.1 Building the graph

Algorithm 1 details how we derive the interactions between users. We start from the list of users of a generic entity in social media; an entity can be, for instance a Facebook page, or a blog's page. We then consider a list of interactions which depends on the type of analyzed entity. This means that interactions that can be derived from a Facebook page are different from those related to a blog page. This is due to the fact that different actions can be performed on different entities; on a Facebook page a user can post, comment or like UGC published by other users. On a blog's page instead, users can only comment other UGC. Examples of actions are then, "like a post" on Facebook or "comment a post" on a blog.

Users involved in possible interactions are defined as source user, for the users performing the action, and target for the user involved in the action itself. Interactions we consider can be single if only two users are involved, or multiple if it is possible to identify a single source user and more than one target user. Given these two lists, we start to analyze crawled data considering every single user involved. The user under analysis is considered as the source of the interaction. We analyze all actions performed by the source users and, for each action, we set the correspondent interactions from the original `listOfInteractions`. If the interaction is single we get the target user of the action and then we set an indirect link. Otherwise, if the interaction deriving from the action is multiple, we consider all others users involved and we set a link for each of these users.

It is important to underline that this algorithm is structured to be entity-independent. In this way, different types of entities, belonging to different types of social media, can be analyzed. Furthermore, the multi-entity nature of the algorithm enables the analysis of more than one entity in a single "round" of analysis.

Algorithm 1 Determine users interaction

```
for all entity e involved in the analysis do
  getEntityUsers
  getListOfInteractions
  getAllAction
  for all user u on the entity do
    for all action a performed by u do
      getRelatedInteraction i
      if i is single then
        get the target user q
        set a link u → q
      else
        get other user j involved in a
        for all other user j do
          set the link u ↔ q
        end for
      end if
    end for
  end for
end for
```

This is important especially regarding to intelligence activities. Indeed, LEAs involved in fighting organised crime, have to deal with groups of people that interact each other exploiting, for instance, more than one Facebook page, typically related to different geographical area in which the organization is present. Moreover, a multi-entity analysis enables studying the behaviour of users that play a key role in their own group and, typically, act as bridges with other similar groups.

8.2 Analysis of interactions - Facebook case study

8.2.1 Facebook Data

To better explain the application of the above-described algorithm to our case study, a brief introduction to the FB data is necessary.

For the scope of our analysis, we will distinguish two possibles objects involved in interactions; *posts* and *comments*. A post is a UGC, published by the administrator (the user who manages the page) or by a generic user of the page, which can contain text, photos, video, link. A comment is similar to a post but with the difference that it represents a reply to a post.

Three different actions can be performed on these objects; *like*, *comments*, and *mentions*. Indeed, even if the objects and the actions can look similar, from a technical point of view, they need to be treated differently due the associated metadata. Clicking on *like* under a UGC on Facebook is the easy way to let someone know that you enjoy it, without leaving a comment. Just as a comment, the fact that you liked it is noted beneath the item. A *comment* is a more explicit way to interact (agree or disagree) with the author of a published UGC. A *mention* is a particular type of tag. These features allow users to create a link between a user and a UGC, such as a photo. The mention is a tag performed on a post or a comment. As an example, Figure 8.1 shows a Facebook page with related objects and actions.

In our analysis we did not include the *share* action because Facebook does not provide information about who shares a UGC. Only numerical data (how many users) about sharing activities are provided by Facebook APIs.

Typically social media that deal with intimate and potentially sensitive data tend to implement rather strict privacy policies. This is also true in the Facebook case. It is possible to summarize these restrictions as follows ([43]):

- Credentials of a Facebook user are mandatory to access data. For example, detailed user data can generally only be extracted from accounts a user is friend with.
- Users' privacy settings play a key role. If one user excludes another from seeing certain elements on his/her profile, an application operating with the latter's credentials will also be blocked from accessing those elements.
- Every application is required to explicitly ask for permission to access different data elements. A pop up, appearing the first time a users access the application, will ask for these permissions.
- Elements visible to the users might not be available through the APIs.

Table 8.1 resume all the possible actions on a Facebook page.

Action/Object	**Post**	**Comment**
Like	like a post	like a comment
Comment	comment on a post	comment on a comment
Mention	mention in a post	mention in a comment

Table 8.1: Possible actions on a FB page

Figure 8.1: Example of Facebook page with related actions

8.2.2 Interactions on Facebook

In Facebook, users can interact in many ways (e.g., messaging, applications, photo uploads, and chat). While posting on a *wall* is one of the most popular methods of user interactions, we are unable to determine if it is representative of other forms of interactions.

Description	Type	FB object	Action
u likes a post published by q	single	post	like
u comments a post published by q	single	post	comment
u is mentioned by q in her post	single	post	mention
u and q comment the *same* post	double	post	comment
u and q are mentioned in the *same* post	double	post	mention
u likes a comment published by q	single	comment	like
u comments a comment published by q	single	comment	comment
u is mentioned by q in her comment	single	comment	mention
u and q comment the *same* comment	double	comment	comment
u and q are mentioned in the *same* comment	double	comment	mention
u likes a comment on a post published by q	single	post	like
u comment a comment on a post published by q	single	post	like

Table 8.2: Possible interactions on a Facebook page

Starting from the above-listed general actions on a FB page, we present in Table 8.2 the possible interactions between users, grouped by objects involved. It is possible to highlight that interactions derived from actions on posts and on comments are very similar. We can also distinguish two types of interactions: those involving the publisher of the UGC and the users performing an action on this UGC, and those involving users performing the same actions on the same UGC. This second type of interactions is very frequent and, especially from a computational point of view, causes a lot of noise in the resulting graph.

According to these discussions, the algorithm to build the interaction graph of a Facebook page is detailed as follows: The same algorithm can be adopted for

Algorithm 2 Building interactions graph of a Facebook page

```
getPageUsers
getListOfPossibleInteractions
getAllActionOfThePage
for all user u on the page do
  for all action a performed by u do
    if a is monodirectional then
      get the target user q
      set the link u → q
    else
      get other user j involved in a
      for all other user j do
        set the link u ↔ q
      end for
    end if
  end for
end for
```

the analysis of a Facebook group or event.

The network that emerges from this process is composed by nodes representing users and links summarizing the interactions. At this stage, in which interactions are not distinguished by some parameters (e.g. their weights), the resulting graph is very dense.

8.3 Weighted Interaction Graph

In order to perform an in depth study on the interactions graph on SM, we also consider the fact that different interactions, especially in the context of social networks,

are characterized by a different level of *strength*. For instance, let us consider two users who interact to each other using the same Facebook page or group. When a user "A" mentions user "B" in a comment on in a post, both "A" and "B" know each other (maybe virtually): between "A" and "B" there is an interaction direct and explicit. We can then infer that this relation is stronger than the one which exists between two users who like the same post. To deal with this aspect we built a weighting system for the interactions graph inferred from Facebook data. For the best of our knowledge, as emerged from the state of the art in Chapter 5, this is the first study on weighted interaction in SM.

The main reasons, from our point of view, of the introduction of a weighting system for interactions graph analysis are two: a theoretical and a practical one. The theoretical reason has been already anticipated: among all the possible interactions involving two generic users of a SN, some are stronger than others. This means that, especially if we think about an application to an intelligence activity, some interactions are meaningful and express an interesting relation whereas some others are meaningless. This means also that including meaningless interactions on the analysis process results into a noise in the output. We will refer to the noise as the computational overhead in front of meaningless derived information caused by some interactions. The practical reason is strictly related to this point; analyzing the noise represented by less stronger interactions is computational expensive, especially due to the fact that, typically, these interactions are frequent.

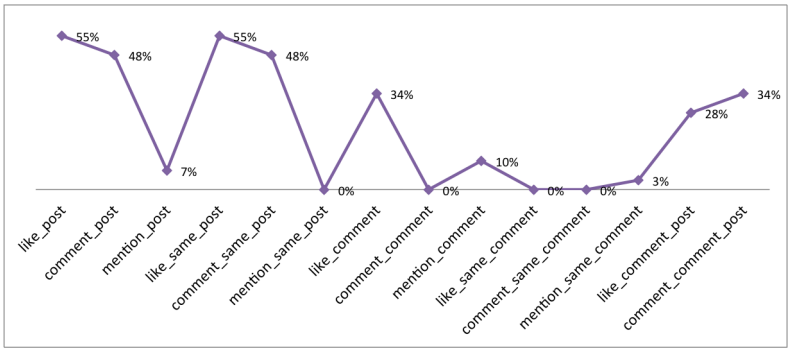


Figure 8.2: Percentage of interactions in the sample

The idea at the base of the proposed weighting system is strictly related to the noise concept and to the fact that rare information are more precious than recurring

ones. We approached the problem from two different perspectives; we conducted a frequency study of different interactions in order to determine which are the rare ones. In addition to this, we also surveyed a group of Facebook users asking them to rank, in a priority order, from the stronger to weaker, possible interactions.

For the frequency study we used a pool of Facebook pages suggested by LEA involved in CAPER project (see Section 4.3). To obtain these pages, LEAs operators were asked to insert, in the crawling module we implemented, some keywords that can be meaningful for an intelligence analysis.

In Figure 8.2 we reported the percentage of different types of interactions and our sample of Facebook pages. It is possible to notice that, comment_comment interactions, related to a situation in which a user "A" comments a comment published by another user "B", is not present; this because the feature allowing users to comment other comments was recently introduced by Facebook and, a lot of users are not yet aware of this possibility.

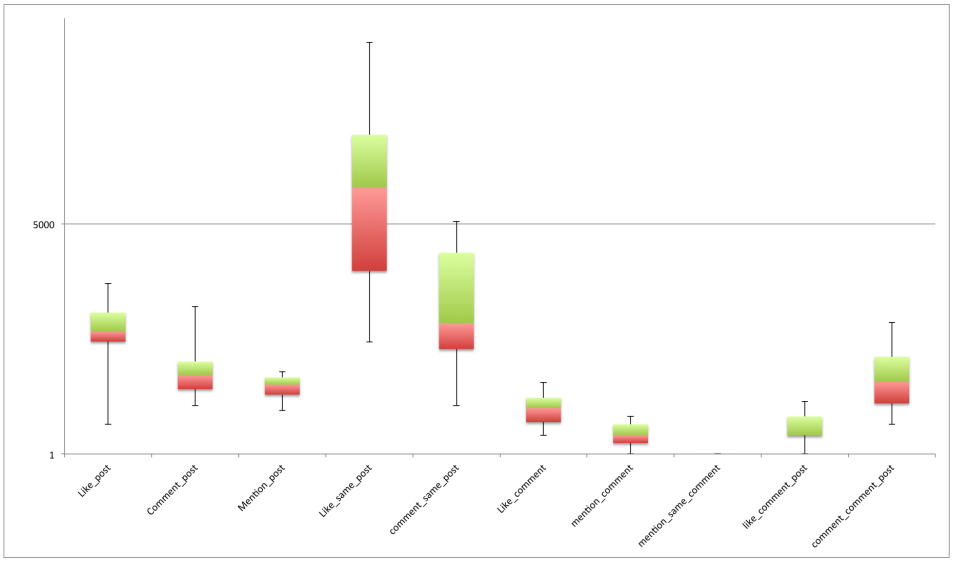


Figure 8.3: Distribution of the frequencies of different interactions

We then studied how much every single interaction was repeated in our dataset. In Figure 8.3 we plotted the frequency of every single interaction we considered. As shown in the graph, more frequent interactions are those related to actions performed on the same Facebook object. For example, likes on the same

post (like_same_post) and comments on the same post (comment_same_post) are the most frequent.

In addition to this, these interactions are less significant; the higher are the number of users involved in the entity, the fewer is the probability that two users who like the same post are interacting each other. In contrast, more "direct" interactions, deriving from actions such as mentioning a specific user on a comment (mention_comment) or in a post (mention_post) are less frequent that others.

Most popular pages on Facebook can involve millions of users; for instance, the page of Rihanna[1] is one of the most popular pages (after Facebook and Facebook for every phone official pages) with 85,501,503 of users involved and a daily grown of +22,296%.

For example, Figure 8.4 shows one of the Facebook pages of the Anonymous organization. The Figure refers to a post addressed to President Obama. Looking at the number of users liking this post it is clear how less significant can be a relation between two these users. The presence of an high number of less significant interactions can be counter-productive in both analytics and computational aspects. The analysis conducted on these interactions can produce weak results, not relevant for intelligence activities. From the computational point of view, calculating interactions between an higher number of users is computationally heavy.

To overcome the problem of the noise produced by less significant interactions we decided to weight each interactions considering following aspects:

- The mean frequency of every single interaction.
- The number of pages under analysis.

Algorithm 3 describes the algorithm we use to determine the weight of an interaction in a group. Starting from this we defined the weight of a single interaction considering:

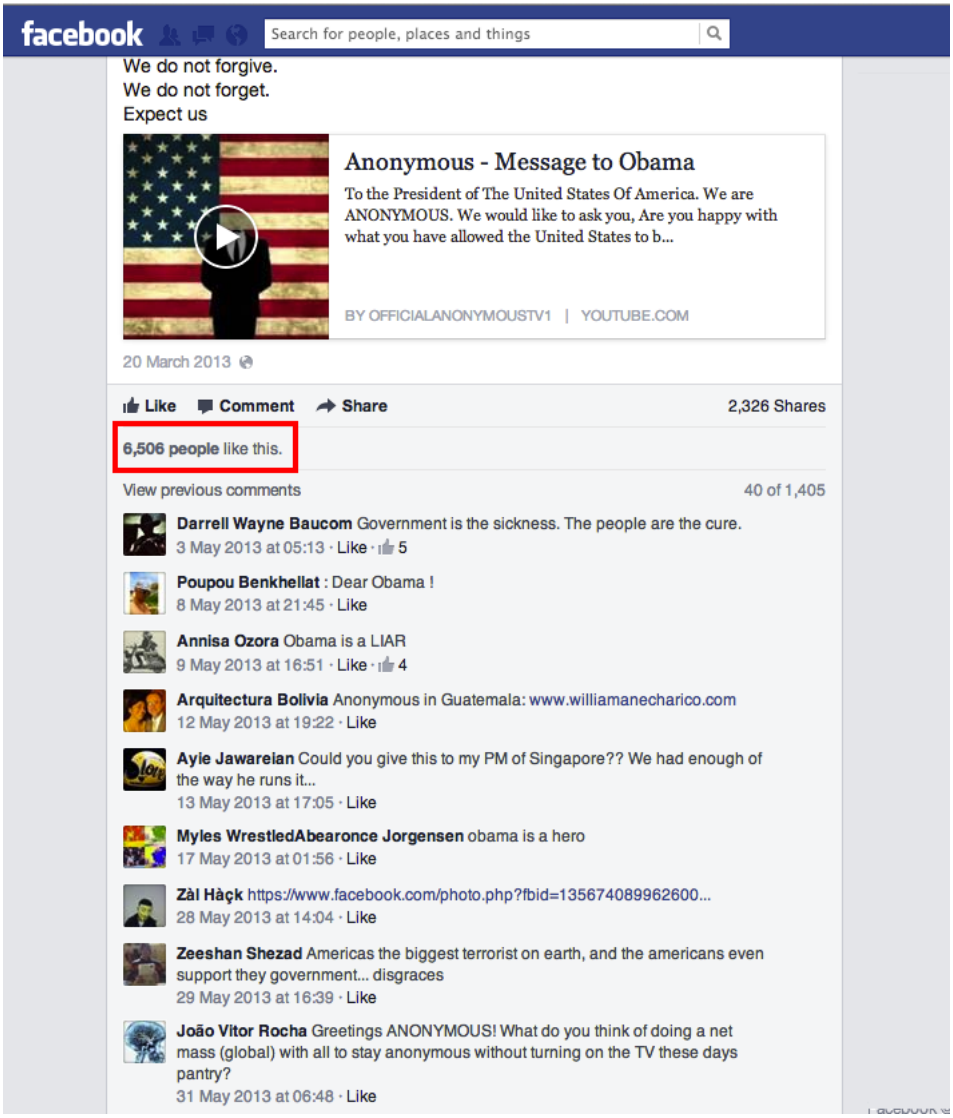
In order to allow a multi-entity approach, we first consider how many entities (pages, groups, events) are analyzed. For each entity under analysis we consider the same list of possible interactions. These interactions were derived as stated in Chapter 8. We consider, then, every single interaction and we count the number of times that it is present in each entity. Given

$$Fi(1)$$

the number of times in which the interaction i appears in entity 1, and considering a dataset composed of n entities, we have

[1] http://www.pagedatapro.com/pages/leaderboard/fc/fan_count

Figure 8.4: Example of the number of likes on a Facebook page



Algorithm 3 Determine the weight of interactions

```
get a list of N entities to be analyzed
for all page p in the list do
   get a list of interactions
   for all interaction i do
      count the number Fi of occurrences of i
   end for
end for
Calculate the mean of Fi in the all list of entities
Calculate Ftot
Calculate the weight of interactions i Wi
```

$$Fm_i = \frac{Fi(1) + Fi(2) + ... + Fi(n)}{n} \tag{8.1}$$

We then calculate the sum of each present on the entity. Assuming that the interactions we consider are m:

$$Ftot = Fm_1 + ... + Fm_m \tag{8.2}$$

Finally we calculate how every single interaction is "important" with respect to others interactions, by using the following formula:

$$1 - \frac{Fm_i}{Ftot} \tag{8.3}$$

Using this formula we then consider, in the weight of a single interaction, following two aspects

- How much is frequent an interaction compared to others appearing in the same entity.
- Most frequent interactions are weaker than rare ones.

Once the weight of every single interaction is calculated, we need to include all these information in the interactions graph.

To do this we summarize all the interactions occurring between each pairs of users and, consequently, we assign a weight to this aggregate interactions calculated as the sum of the weights of the all interactions.

Algorithm 4 describes the procedure we use to build the weighted interactions graph.

The resulting graph will be composed by single pairs of users of the analyzed entity linked. Links between two users will represent all the interactions occurred during the observation period.

Algorithm 4 Determine aggregate interactions

```
get a pair of users in a page
interactionWeight = 0
for all interactions between the pair do
    get the corresponding weight
    sum the weight to interactionWeight
end for
```

9

Results

In this Chapter we present the results of the testing of the algorithm that generates the interaction graph. Results are for the *unweighted* version, which considers all the possible interactions at the same level, and for the *weighted*, in which every interaction is instead characterized by a weighting parameter.

To perform our tests we used a Facebook group. We crawled the group's data, including feeds and users, and we applied our algorithm for the extraction of the interactions graph. The obtained network is a *full* (or *complete*) and *unimodal* network. It is a full network because it contains all the users in the group and the connections among them; all the "egos" are treated equally. It is unimodal because it includes one type of vertex, namely users.

The result is a graph $G := (V, E)$ composed of $|V|$ (vertices or nodes) users of the analyzed Facebook group and $|E|$ (edges) different interactions (in the sense of possible actions on Facebook) between these users. We studied both graphs, weighted and unweighted, using standard metrics in network analysis theory. For each graph we used both aggregate and vertex-specific networks metrics. While aggregate metrics describe entire networks, vertex-specific metrics identifies individuals' positions within a network.

As aggregate metrics we considered:

- Maximum geodesic distance.
- Average geodesic distance.
- Graph density.

The maximum geodesic distance, also referred as diameter of a network, is the largest geodesic distance of all, or the distance between the two vertices that are farthest from each other.

The average geodesic distance gives a sense of how "close" community members are from one another. If it is low, as in our study, most people know one another either directly or through a mutual friend. This is confirmed by the fact that we are investigating a group on Facebook composed by people that share same interests and use this space on Facebook to discuss about it. An higher value indicates, on the other hand, that many individuals in the social network do not directly know each other.

Graph density is an aggregate network metric, ranging from 0 to 1, used to describe the level of interconnectedness of the vertices. It is a quantitative way to capture important sociological ideas like cohesion, solidarity, and membership. For our undirected graph, due the fact that all vertices are connected to all others through at least one edge, the graph density is calculated by dividing the number of total edges by the maximum number of possible edges.

Regarding vertex-specific metrics, we focused on centrality measures. The concept of centrality is one of the most important and commonly used for exploring actor roles in social networks. Centrality is a basic but essential measure in social network analysis because enables the identification of which nodes are in the "center" of the network. For the scopes of our research this is very useful because it allow us to pinpoint users that are more interesting for investigation activities. In particular, we focused on following metrics:

- Degree centrality.
- Betweenness centrality.
- Closeness centrality.

Degree centrality is a count of the total number of connections linked to a vertex. The degree centrality of a vertex v is defined as:

$$C_D(v) = \deg(v) \tag{9.1}$$

As stated in [51], this metrics suggests that "central actors must be the most active in the sense that they have the most ties to other actors in the network or graph".

The betweenness centrality, typically used in complex networks, especially in social networks [7, 15] studies, is based on the concept of how far are two vertices (not neighbors) in a network. The distance between nodes who are not neighbors is measured by the smallest number of neighbor-to-neighbor hops from one to the other. The shortest path that we can notice in the network between two nodes is called the *geodesic distance*. According to [4] betweenness centrality focuses on "the share of times that a node *i* needs a node *k* (whose centrality is being measured) in order to reach *j* via the shortest path". In this meaning, *k* can be thought

of as a kind of "bridge" and betweenness centrality as a measure of how much removing *k* would disrupt the connections between other nodes in the network.

The betweenness centrality of a node v is given is calculated as the number of the shortest paths from all vertices to all others that pass through that node; it is given by the equation:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{9.2}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

Closeness centrality indicates the influence of a node on the entire network, capturing the average distance between a vertex and every other vertex in the network. A low closeness centrality means that a person is directly connected or "just a hop away" from most others in the network. In contrast, vertices in very peripheral locations may have high closeness centrality scores; this value indicates the number of hops or connections they need to take to connect to distant others in the network. Closeness can be defined in several ways. It can be defined as the average length of the of the shortest paths from one node to all other nodes [51]. Moreover, the closeness centrality of a vertex is the mean geodesic distance between it and all other vertices reachable from it [40].

In some other cases the inverse of the average distance to others in the network is used as a measure of closeness centrality. In that case, higher values indicate a more central position:

$$C_C(v) = \left[\sum_{j=1}^{g} d(n_i, n_j) \right]^{-1} \tag{9.3}$$

Table 9.1 shows the meanings of the correlation between the three above-mentioned centrality metrics.

To perform the evaluation we asked to the group's members to describe their interactions with other members. To each member were asked the following questions:

1. How often do you interact, using groups facilities with each one of the others members?
2. Rate in a priority order possible interactions.

We then compared the manually designed version with those extracted using our algorithms.

	Low Degree	**Low Betweenness**	**Low Closeness**
High Degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant communication bypasses him/her
High Closeness	Key player tied to important/active alters		Probably multiple paths, ego is near many people
High Betweenness	Ego's few ties are crucial for network flow	Very rare cell	

Table 9.1: Comparison between centrality measures

During the observation period active users was the 64% of the total member of the groups. We conducted our tests only on these users because we are interested in interaction between active users instead of their friendship relations.

In Figure **??** is represented the interactions graph obtained by our algorithm. The version in this Figure is the unweighted one: all interactions are considered at the same level.

Vertices are displayed according to their vertex-specific metrics. The vertex color indicates the betweenness centrality and range from blue to orange: nodes with lower betweenness centrality are in blue, while higher value are indicated in orange. The dimension of the nodes is used to represent the degree centrality. Bigger nodes have an higher degree centrality, while smaller ones have fewer connections between other nodes.

Graph Type	Directed
Vertices	59
Unique Edges	291
Edges With Duplicates	1011
Total Edges	1302
Connected Components	1
Maximum Vertices in a Connected Component	59
Maximum Edges in a Connected Component	1302
Maximum Geodesic Distance (Diameter)	4
Average Geodesic Distance	1,79
Graph Density	0,14

Table 9.2: General graph metrics of the weighted graph

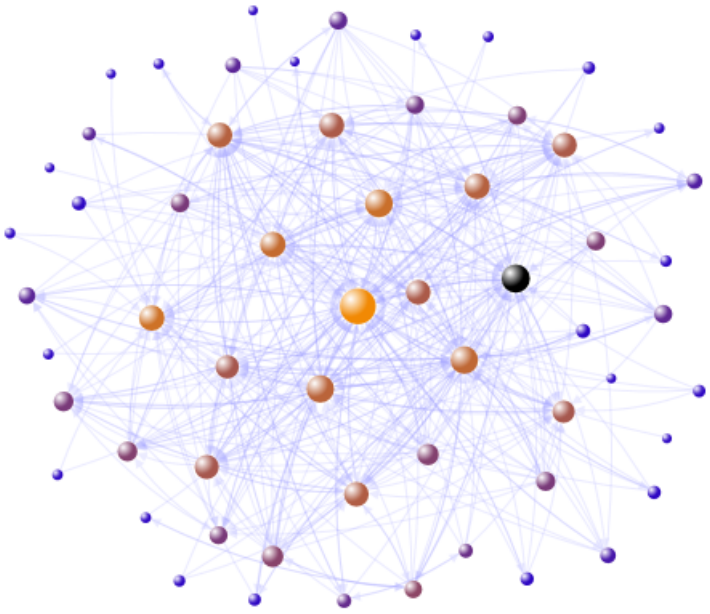


Figure 9.1: Unweighted graph automatically extracted

Table 9.2 resumes the overall metrics for the unweighted version of the graph. The metrics are calculated using a tool for social network analysis: NodeXL [27]. NodeXL is used also to generate the graph images.

All the users in the network compose a unique connected component; this means that there are no clusters of vertices that are connected to each other but separate from other vertices in the graph. It is interesting to note that, by considering all interactions at the same level, we have an high number of duplicated edges. Duplicate vertex pairs represent situation like, for example, the reply/response flow between two users on a post.

The geodesic distance is the length of the shortest path between two users. Considering that edges represent roads and nodes are cities, the geodesic distance would be the number of roads someone must take to get from city to another, assuming that the person is traveling on the shortest path possible. The maximum geodesic distance, or diameter of a network, is calculated as the largest geodesic distance of all, or, in other words as the distance between the two vertices that are farthest from each other.

The average of all geodesic distances indicates how "close" group members are from one another. An higher value of this metrics means that many individuals in the social network do not directly know each other. People may be connected through a friend of a friend of a friend of a friend, but not through short paths. A low value indicates, on the other hand, that most people know each other either directly or through a mutual friend. In our study people belonging to the analyzed group directly knew each other: this is confirmed by the obtained value of the average geodesic distance.

The graph density is a value, ranging between 0 and 1 indicating how interconnected the vertices are in the network. A more dense graph (e.g., 0.6) would include more total edges for the same number of vertices.

On the other hand, Figure 9.2 represents the weighted version of the interaction graph. The centrality metrics are plotted in this case using the same parameters (color, size) that the unweighted version. In addition to this, we plotted the strength of the connection between users using the size of the links.

Pairs of users connected by bigger links are those with a bigger number of interactions. Moreover, according to the weighting formula, the size of the link is influenced also by the "importance" of the single interaction in the context of the analyzed group.

To better underline the difference between links, Figure 9.3 shows the top 20 connections, in terms of weight, in the obtained graph.

As expected, users in this graph are those with higher values of centrality metrics.

Graph Type	Directed
Vertices	59
Unique Edges	485
Edges With Duplicates	0
Total Edges	485
Connected Components	1
Maximum Vertices in a Connected Component	59
Maximum Edges in a Connected Component	485
Maximum Geodesic Distance (Diameter)	4
Average Geodesic Distance	1,79
Graph Density	0,14

Table 9.3: General graph metrics of the weighted graph

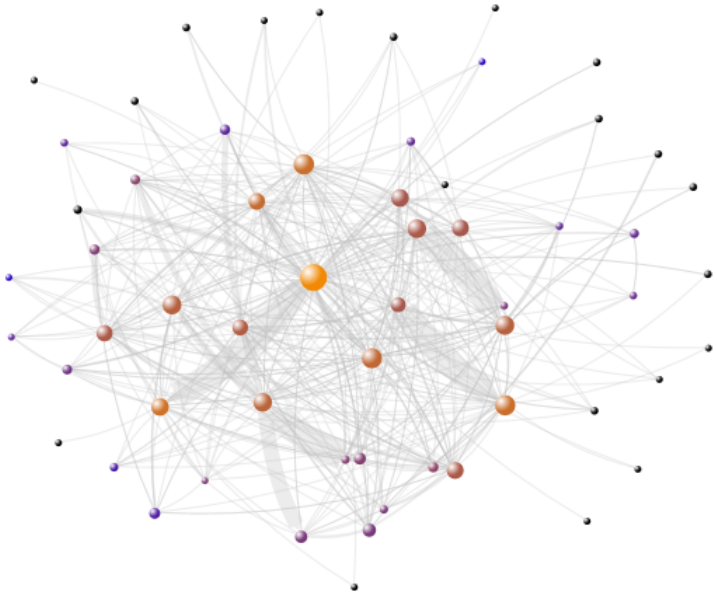


Figure 9.2: Weighted Interaction graph - automatically extracted

Figure 9.3: Top 20 connections in the interactions graph

Table 9.3 contains general graph metrics for the weighted graph. The numbers of nodes composing the graph is 59, connected by 485 unique edges. As expected, the number of edges with duplicates is 0. This is due to the fact that, in the weighted version of the graph, we summarize all the interactions between users using all related weights. As in the unweighted version, all the users are connected in a unique connected component; there are no nodes acting as bridge between one community to another. As expected, the diameter of the network and the average geodesic distance are the same: considering interactions in an aggregate way does not affect the closeness of users.

Figures 9.7, 9.4, 9.5, and 9.6 show frequency charts for the considered vertex-specific metrics. These charts plot the number of users with similar value of the considered centrality metrics. For example, in the in-degree chart, it is possible to notice that the majority of users have a small number of incident edges (so are targets in the interactions). On the other hand, the out-degree chart shows that all the users tend to interact (as source users) with other users and, as a consequence, they have similar value of out-degree.

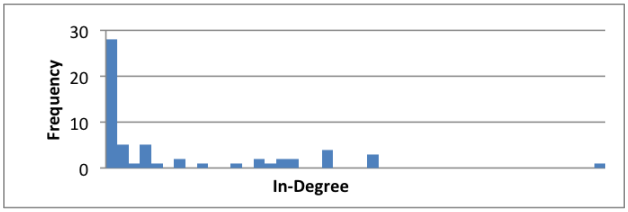


Figure 9.4: In-Degree centrality - Weighted interactions graph

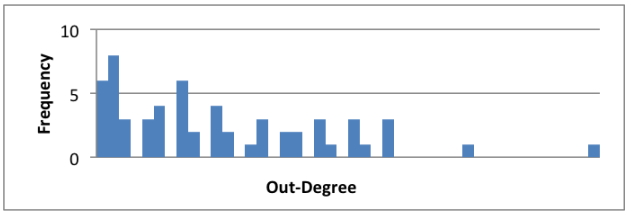


Figure 9.5: In-Degree centrality - Weighted interactions graph

We now discuss the comparison between the result produced by our algorithms with the graphs manually extracted from the sample group. We derived

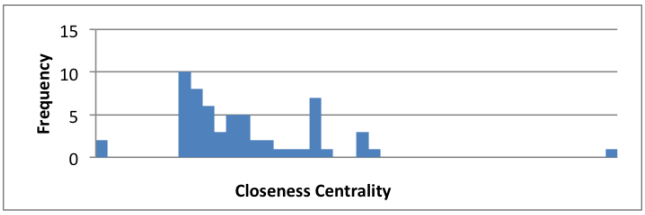


Figure 9.6: In-Degree centrality - Weighted interactions graph

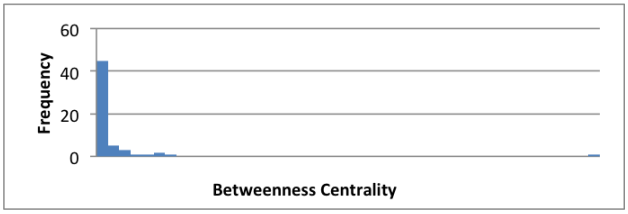


Figure 9.7: In-Degree centrality - Weighted interactions graph

interactions from the answers provided by the groups users during the survey. Figure 9.8 shows on right the graph derived from the application of interaction extraction algorithm to sample data extracted with our framework for the managing of social media pipeline; on left, instead the figure shows the graph manually extracted from the answers.

It is mandatory to underline that, among the 64% of active users of the group we analyzed with the algorithm, the 41% provided answers useful for the evaluation. With useful answers we refer to answers which are complete, providing information about all other users. To deal with this problem and a more accurate comparison we removed both users who were active but did not provide complete answers to the survey and users who answered to the survey but were not active in the crawling period.

Figure 9.8 depicts the interactions graph extracted from the survey of the group's users. We used the same parameters of the automatically extracted graph; the color of the nodes represents the betweeness centrality and the size the degree centrality.

As in the automatic extracted graph, few users have an higher value of degree so they act as hub for the interactions in the group.

Table 9.5 lists graph metrics for the interactions graph extracted from the survey results. The difference between the number of total edges of the automatically extracted graph and the one resulting by the survey can be due to the fact that we

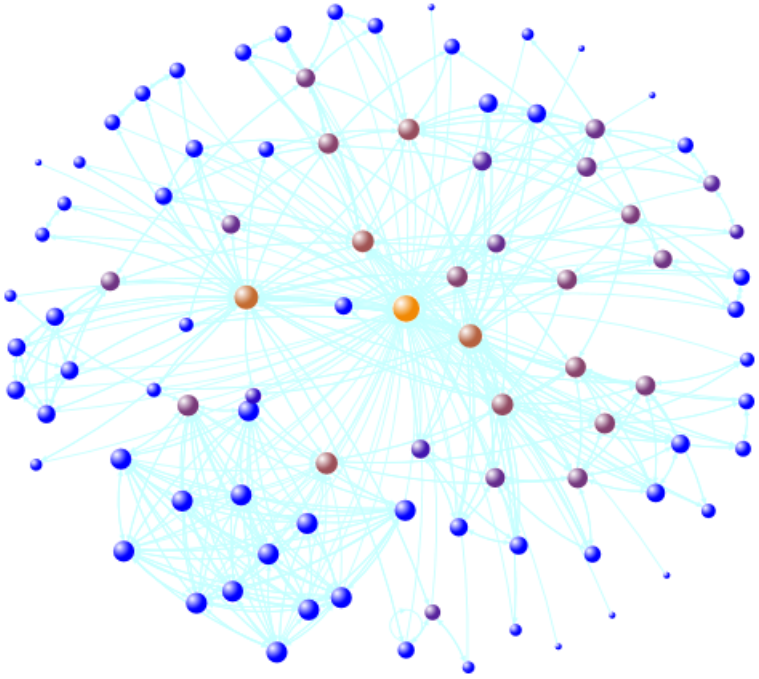


Figure 9.8: Unweighted interaction graph - Derived from survey

Graph Type	Directed
Vertices	59
Unique Edges	350
Edges With Duplicates	1035
Total Edges	1385
Connected Components	1
Maximum Vertices in a Connected Component	59
Maximum Edges in a Connected Component	1385
Maximum Geodesic Distance (Diameter)	3
Average Geodesic Distance	1,61
Graph Density	0,19

Table 9.4: General graph metrics of the unweighted graph extracted from the survey

removed, in our algorithm, the most frequent interactions: "like the same post" and "comment the same post". Regarding other metrics results have shown that the extracted version of the interaction graph is very similar to the real one.

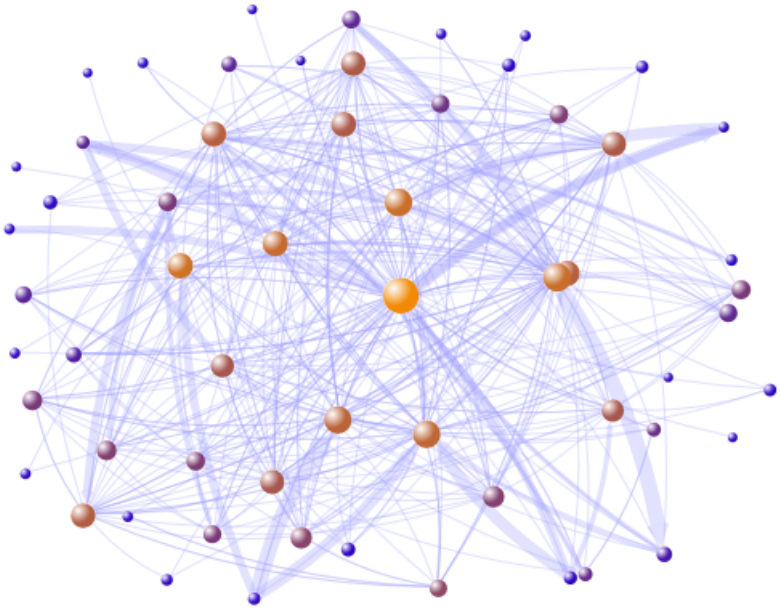


Figure 9.9: Weighted interactions graph - Extracted from the survey

Figure 9.9 represents the weighted version of the graph. In this case the weight of the aggregate interactions is derived from the answers provided in the survey. In particular, the weights of the interactions were derived from the priority order provided by users. Starting from the weights obtained, we applied the algorithm to derive the weighted interactions graph.

Similarly to what obtained from the automatic extracted graph, in this case general graph metrics of the weighted version of the graph are similar to those of the unweighted version. The aggregation of the interactions results in the same number of unique edges that in the automatic extracted version.

Graph Type	Directed
Vertices	59
Unique Edges	485
Edges With Duplicates	0
Total Edges	485
Connected Components	1
Maximum Vertices in a Connected Component	59
Maximum Edges in a Connected Component	485
Maximum Geodesic Distance (Diameter)	3
Average Geodesic Distance	1,61
Graph Density	0,19

Table 9.5: General graph metrics of the unweighted graph extracted from the survey

Figures 9.13, 9.10, 9.11, and 9.12 show frequency charts for the considered vertex-specific metrics. These charts plot the number of users sharing the same value of the considered centrality metrics.

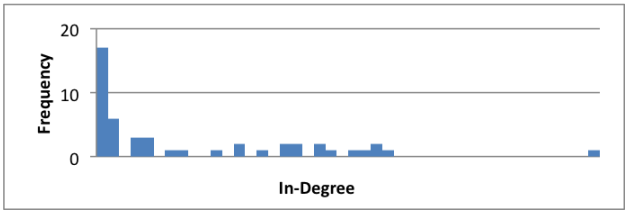


Figure 9.10: In-Degree centrality - Weighted interactions graph

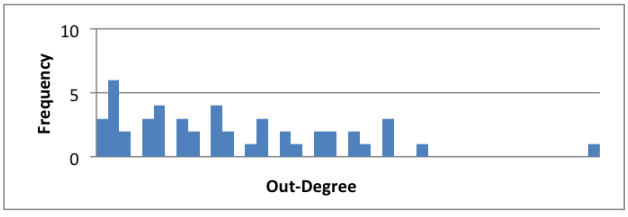


Figure 9.11: In-Degree centrality - Weighted interactions graph

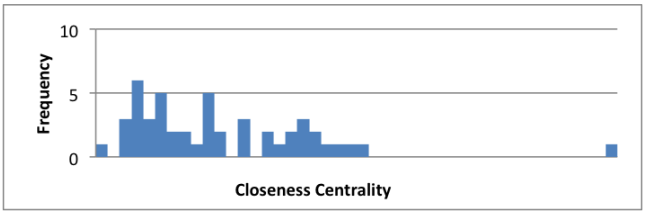


Figure 9.12: In-Degree centrality - Weighted interactions graph

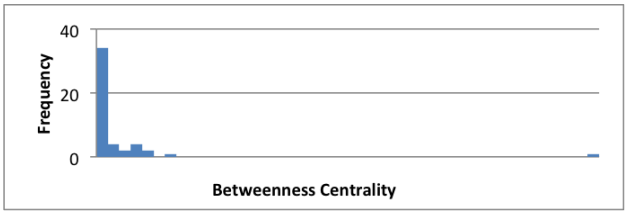


Figure 9.13: In-Degree centrality - Weighted interactions graph

10

Conclusions

Global criminals are now sophisticated managers of technology using ICT tools in a careful and planned way. Central networked intelligence and coordinated knowledge are fundamental assets available also to non-terrorist criminal organizations. Homeland security and intelligence analysis communities use open source information, which comes from publicly available sources. They rely on valuable instruments to perform intelligence activities and investigations, the so-called Open Source Intelligence (OSINT).

Due to the widespread use of the Internet, one of the most important source of OSINT is the Web and, in particular Social Media. Social media contains an enriched set of data and metadata that can be useful in the intelligence activities. In order to collect, monitor, analyze, summarize, and visualize social media data tools and frameworks have been developed in the so-called field of Social Media Analytics (SMA). By exploiting technology, frameworks, and tool sets from SMA, Social Media Intelligence (SOCMINT) aims to derive actionable information from SM for applications that can benefit from the “wisdom of crowds” through the Web. The term refers to a wide range of applications, techniques and capabilities exploiting social media data in intelligence.

Although SOCMINT is used not only in the investigation and crime fighting environment, but also for marketing and industrial scopes, nowadays a large amount of data processing is still conducted manually; this results in a waste of both human and time resources. There are software services to perform automatic processing, but the functionality and degree of automation are still immature and limited, especially considering the vast amount of data today available on Social Media. As a result there is a need of instruments helping in SOCMINT activities from different perspectives; from the collecting and crawling phases to the analysis and visual-

ization processes. Possible solutions have to take into account several limitations both ethical and technological.

This thesis provided contributes in the field of Social Media Intelligence, by answering following questions: i) How a Law Enforcement Agency (LEA) can successfully exploits social media data? ii) Which types of analysis could be useful in an intelligence environment?

The problems have been approached in a top-down fashion: first we studied the problem of how to usefully exploit and deal with social media data in intelligence activities and propose an approach to address the issue. Then we moved on a more specific area: the problem of how to analyse social media data in order to provide to the LEAs useful results and provide, specifically, an analysis of Facebook data. As third main contribute, we improved the analysis proposed by studying a method that takes into account the nature of Facebook data.

The problem of black hole in the dataset to be analyzed was studied and a framework for the managing of social media pipeline has been proposed. A social media pipeline is a workflow starting with the retrieving of raw data from SM sources and ending with analyzed information. The framework, due to his modular nature, is able to deal with different types of social media and can analyze entities (such as pages, groups, etc..) in parallel. The framework can manage the collection, the processing and the analysis of raw data collected from social media.

Regarding the analysis of social media, we approached the privacy related problems, dealing with friendship networks, as typically happens in social media data. Due to the privacy settings managed by users, a lot of interesting information (e.g. including the list of friends) are often not public. To overcome this problem, we chose to use data in spaces in SM that are "public-by-design"; pages or groups on Facebook, comments on public posts on blogs, etc.. In this way two problems deriving from privacy settings can be managed: the presence of black holes in the dataset and the compliance to ethical and privacy requirements. Moreover, even if the friendship network is a good indicator of different social aspects, people interacting each other on some topic or discussion, including those related to illegal activities, are often not "friends". Consequently their relations do not emerge, for example, from analysis of the ego networks.

To improve the resulting interaction graph we also designed a weighting system that assign a weight at each interaction on a entity depending on i) how much is frequent an interaction compared to others appearing in the same entity; ii) the fact that most frequent interactions are weaker than rare ones. The evaluation of the results was conducted using a real Facebook group. Users were surveyed about

how they conduct their interaction in the group and which are, in their opinion, the most valuable interactions.

Results show that looking at interactions, instead at friendship, allowed us to discover more useful information. The interaction graph gives us a more complete picture of how users communicate each other. Moreover, our results confirm the fact that active users are key players in a network. The interaction graph extracted by our algorithm and the one manually retrieved from the answers of the users are the very similar. In addition to information available from the manually extracted graph, the graph resulting from the algorithm provides additional interactions that, otherwise, were not visible.

First, we must take into account that types of data manipulated in these activities are strictly dependent on privacy settings; this means that, for instance, an analysis of a friendship graph could be not possible because of the lack of the (whole) dataset. Furthermore, due to the basic idea of the OSINT paradigm, only public information can be used in the process; therefore it should be avoided a dependency of the dataset on privacy settings of the analyzed social media. Moreover, looking at the technological aspects of SOCMINT process, there are great challenges due to the huge quantity of data involved; some of them include data storage and analysis. Summarizing, our main contributes are the following: i) the design of a general framework for the management of social media data; ii) the design of an algorithm to extract interactions graphs from Facebook data; iii) a weighting system for the interaction graph that takes into account different types of possible interactions and provides more accurate results.

The proposed framework deals with two important aspects of a social media pipeline. We defined this pipeline as a workflow beginning with the retrieving of raw data from social media sources and ending with analyzed information. In order to easily deal with different types of social media, the framework has been designed with a modular architecture. This enables the extension of the framework itself in a very easy way.

References

1. Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.
2. Jamie Bartlett and Carl Miller. The state of the art: a literature review of social media intelligence capabilities for counter-terrorism. *http://www.demos.co.uk/*, 2013.
3. Clive Best. Open source intelligence. *Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and Their Applications to Security*, 19:331, 2008.
4. Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.
5. Stephen P Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1):21–34, 2006.
6. Stephen P Borgattr. The key player problem. In *Dynamic social network modeling and analysis: Workshop summary and papers*, page 241. National Academies Press, 2003.
7. Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.
8. John Carr. New approaches to dealing with online child pornography. In *Cybersecurity Summit (WCS), 2011 Second Worldwide*, pages 1–3. IEEE, 2011.
9. Duen Horng Chau, Shashank Pandit, Samuel Wang, and Christos Faloutsos. Parallel crawling for online social networks. In *Proceedings of the 16th international conference on World Wide Web*, pages 1283–1284. ACM, 2007.
10. Kim-Kwang Raymond Choo. Organised crime groups in cyberspace: a typology. *Trends in organized crime*, 11(3):270–295, 2008.
11. Kim-Kwang Raymond Choo and Russell G Smith. Criminal exploitation of online systems by organised crime groups. *Asian journal of criminology*, 3(1):37–59, 2008.
12. Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 57–70. ACM, 2008.
13. Robert M Clark. *Intelligence analysis: a target-centric approach*. CQ press Washington, DC, 2004.
14. Scott Cunningham and Todd D Kendall. Prostitution 2.0: The changing face of sex work. *Journal of Urban Economics*, 69(3):273–287, 2011.

15. Alfredo Cuzzocrea, Alexis Papadimitriou, Dimitrios Katsaros, and Yannis Manolopoulos. Edge betweenness centrality: A novel algorithm for qos-based topology control over wireless sensor networks. *Journal of Network and Computer Applications*, 35(4):1210–1217, 2012.
16. Alex De Joode. Effective corporate security and cybercrime. *Network Security*, 2011(9):16–18, 2011.
17. Brice De Ruyver and Tom Vander Beken. *Measuring organised crime in Belgium: a risk-based methodology*. Maklu, 2000.
18. Ratan Dey, Zubin Jelveh, and Keith Ross. Facebook users have become much more private: A large-scale study. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 346–352. IEEE, 2012.
19. Judith Donath and Danah Boyd. Public displays of connection. *bt technology Journal*, 22(4):71–82, 2004.
20. Paul AC Duijn and Peter PHM Klerks. Social network analysis applied to criminal networks: Recent developments in dutch law enforcement. In *Networks and Network Analysis for Defence and Security*, pages 121–159. Springer, 2014.
21. Nicole B Ellison, Charles Steinfield, and Cliff Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.
22. Scott Garriss, Michael Kaminsky, Michael J Freedman, Brad Karp, David Mazières, and Haifeng Yu. Re: Reliable email. In *NSDI*, volume 6, pages 22–22, 2006.
23. Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.
24. Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.
25. Marc Goodman. What business can learn from organized crime. *Harvard Business Review*, 89(11):27–30, 2011.
26. Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80. ACM, 2005.
27. Derek Hansen, Ben Shneiderman, and Marc A Smith. *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann, 2010.
28. Walter R Harper and Douglas H Harris. The application of link analysis to police intelligence. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 17(2):157–164, 1975.
29. Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.
30. Peter Klerks. The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? recent developments in the netherlands. *Connections*, 24(3):53–65, 2001.
31. Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2006.

32. Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.
33. Frank McCown and Michael L Nelson. What happens when facebook is gone? In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 251–254. ACM, 2009.
34. Don McDowell. *Strategic intelligence: a handbook for practitioners, managers, and users*. Scarecrow Press, 2008.
35. Cathleen McGrath, Jim Blythe, and David Krackhardt. The effect of spatial arrangement on judgments and errors in interpreting graphs. *Social Networks*, 19(3):223–242, 1997.
36. Cathleen McGrath, David Krackhardt, and Jim Blythe. Visualizing complexity in networks: Seeing both the forest and the trees. *Connections*, 25(1):37–47, 2003.
37. Stanley Milgram. The familiar stranger: An aspect of urban anonymity. *The individual in a social world*, pages 51–53, 1977.
38. Alan Mislove, Krishna P Gummadi, and Peter Druschel. Exploiting social networks for internet search. In *5th Workshop on Hot Topics in Networks (HotNets06). Citeseer*, page 79, 2006.
39. Derek O.Callaghan, Derek Greene, Maura Conway, Joe Carthy, and PÃ¡draig Cunningham. An analysis of interactions within and between extreme right communities in social media. 8329:88–107, 2013.
40. Kazuya Okamoto, Wei Chen, and Xiang-Yang Li. Ranking of closeness centrality for large-scale social networks. In *Frontiers in Algorithmics*, pages 186–195. Springer, 2008.
41. David Omand, Jamie Bartlett, and Carl Miller. Introducing social media intelligence (socmint). *Intelligence and National Security*, 27(6):801–823, 2012.
42. Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo De Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179, 2007.
43. Bernhard Rieder. Studying facebook via data extraction: the netvizz application. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 346–355. ACM, 2013.
44. Jose A Scheinkman. Social interactions. *The New Palgrave Dictionary of Economics*, 2, 2008.
45. Daniel M Schwartz and Tony DA Rouselle. Using social network analysis to target criminal networks. *Trends in Organized Crime*, 12(2):188–207, 2009.
46. Robert David Steele. Open source intelligence. *Handbook of intelligence studies*, page 129, 2007.
47. Maksim Tsvetovat and Alexander Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. " O'Reilly Media, Inc.", 2011.
48. Renée C van der Hulst. Introduction to social network analysis (sna) as an investigative tool. *Trends in Organized Crime*, 12(2):101–121, 2009.
49. Graham Vickery and Sacha Wunsch-Vincent. *Participative web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Cooperation and Development (OECD), 2007.
50. Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

51. Stanley Wasserman. *Social network analysis.Methods and applications*, volume 8. Cambridge university press, 1994.
52. Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.
53. Gabriel Weimann. Terror on facebook, twitter, and youtube. *Brown J. World Aff.*, 16:45, 2009.
54. Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.
55. Jennifer Xu and Hsinchun Chen. Untangling criminal networks: A case study. In *Intelligence and Security Informatics*, volume 2665 of *Lecture Notes in Computer Science*, pages 232–248. Springer Berlin Heidelberg, 2003.
56. Jennifer J. Xu and Hsinchun Chen. Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. *Decision Support Systems*, 38(3):473 – 487, 2004.
57. Jennifer J. Xu and Hsinchun Chen. Crimenet explorer: A framework for criminal network knowledge discovery. *ACM Trans. Inf. Syst.*, 23(2):201–226, April 2005.
58. Hsin-Chang Yang and Chung-Hong Lee. Mining open source text documents for intelligence gathering. In *Information Technology in Medicine and Education (ITME), 2012 International Symposium on*, volume 2, pages 969–973, Aug 2012.
59. Shaozhi Ye, Juan Lang, and Felix Wu. Crawling online social graphs. In *Web Conference (APWEB), 2010 12th International Asia-Pacific*, pages 236–242. IEEE, 2010.
60. Haifeng Yu, Michael Kaminsky, Phillip B Gibbons, and Abraham Flaxman. Sybilguard: defending against sybil attacks via social networks. *ACM SIGCOMM Computer Communication Review*, 36(4):267–278, 2006.
61. Aaron Zelin. The state of global jihad online. *New America Foundation, National Security Studies Program Policy*, 2013.
62. Daniel Zeng, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. Social media analytics and intelligence. *Intelligent Systems, IEEE*, 25(6):13–16, 2010.

11

Acknowledgments

I would like to express my gratitude to my supervisors and my Research group, the Web Application for the Future Internet at IIT-CNR, which gave me a constant support to perform this work. Thank you to my colleagues working on the CAPER FP7 project; this project gave us the opportunity to conduct this research activity and keep in touch with a great network of European researchers.

A special thank goes to my friends of "Mezzogiorno di Fuoco" chorus; their data, their surveys and, especially their support was fundamental for the thesis work.

Finally, thank you to my husband and to my wonderful family for the support, the comprehension, and for the joy of sharing my accomplishments.