UNIVERSITÀ DEGLI STUDI DI MILANO

SCUOLA DI DOTTORATO IN

Informatica

DIPARTIMENTO DI

Informatica

CORSO DI DOTTORATO

Informatica

XXVIII° Ciclo

TESI DI DOTTORATO DI RICERCA

Multidimensional Analysis of People's Behavior in Online Social Networks

INF/01

Dottorando:
Azadeh ESFANDYARI

Relatore:
Prof. Gian Paolo ROSSI

Correlatore:
Prof. Sabrina Tiziana GAITO

Coordinatore del Dottorato:
Prof. Paolo BOLDI

Anno Accademico 2015/2016

To my dear husband Ehsan

Contents

1 Introduction 1

1.1 People's behavior across social media 3

1.2 User identification across online social networks 6

1.3 The effect of offline sociality on online interaction 7

2 People's behavior across social media 9

2.1 Introduction 9

2.2 The related works 13

2.3 Methodology 16

2.3.1 Data collection methodology 17

2.3.2 Dataset characteristics and representativity 19

2.4 The Usage of multiple social sites 21

2.5 Centrality and activity correlation across social sites 24

2.5.1 A case study on 4 social media. 29

2.6 Temporal patterns in the posting activity 31

2.7 Patterns in username usage. 34

2.8 Conclusion 35

3 User Identification across Online Social Networks 39

3.1 Related work 40

3.1.1 Username based identification 41

3.1.2 Profile based identification 42

3.1.3 Network and content based identification 42

3.2 Dataset 44

3.2.1 Username composition 46

3.3 Methodology . . . 48
3.3.1 Problem definition . . . 48
3.3.2 Feature extraction . . . 49
3.3.3 Training and test sets . . . 50
3.4 Evaluation . . . 55
3.4.1 Finding candidate users . . . 57
3.4.2 Comparison with existing algorithms . . . 59
3.4.3 Features Importance Analysis . . . 61
3.5 Conclusion . . . 62
4 Effect of offline sociality on online interactions . . . 63
4.1 Dataset . . . 64
4.2 Results . . . 66
4.2.1 Communicability . . . 66
4.2.2 Clustering coefficient . . . 68
4.2.3 Degree . . . 70
4.3 Conclusion . . . 73
5 Conclusion . . . 75
References . . . 79

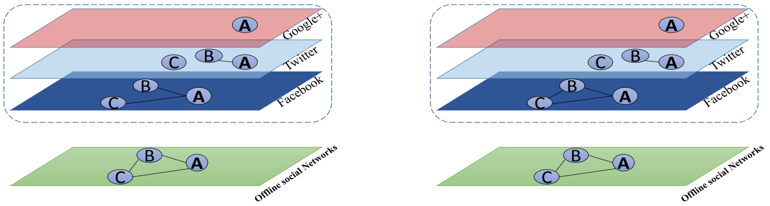
Chapter 1

Introduction

The ever increasing popularity of Online Social Networks (OSNs) is evidenced by the huge number of users who are turning to Facebook,Twitter and other social networks. Unlike the traditional Web, which is largely organized by content, the users are first-class entities in online social networks. The users join a network, publish their own content, and create links to other users in the network. This basic user-to-user link structure facilitates the online interaction by providing mechanisms for organizing both real-world and virtual contacts, for finding other users with similar interests, and for locating content and knowledge contributed or endorsed by friends. The rapid growth of these online social networks provides a unique chance to study and understand the online behavior of the people.

The online behavior of the people is influenced by different factors derived from their real (i.e., offline) and virtual (i.e., online) life. For instance, the friendship acceptance in an online social network might depend on the degree of acquaintance, if the people have already met in some physical places or if they share the same interests.

Studying people's behavior becomes more complicated due to the fact that people in their online life have access to a wide portfolio of social platforms which allow them to differentiate their interests and convey diverse contents. As a result, a user can be registered to many social media and interacts with her/his friends' circles through different channels. For instance, a user can share her/his photos on the Instagram, organize an event on the Facebook and maintain the relationships with workmates on the LinkedIn. Generally speaking, we can gain more knowledge about human behavior by assuming a stratified reality where individuals act as bridges cutting across the various levels and where their decisions are based on information coming from each dimension.



(a) Scenario at time t (b) Scenario at time $t + \Delta t$

Fig. 1.1. The reference scenario of the thesis. At the first level people interact face-to-face and in online level people may registered in several online social media. From time t to time $t + \Delta t$ the online interactions of B and C change and a new link between them in Facebook layer is established maybe due to their offline interactions in offline level.

The reference scenario of this thesis is illustrated in Fig1.1, where we consider two main dimensions: offline and online. Online dimension includes several layers that each one of which represents the social network of a social media. People in offline plane have the real world interactions, while some of these people also have online friendships in several layers of the online dimension. Fig1.1 also shows how real-world interactions in the offline dimension results in interactions in some layer of online dimension. A, B and C have registered in both Facebook and Twitter, whereas A is the only user registered in Google+. It can be noted that the offline interaction of B and C changes their online interactions and a new link between them in Facebook layer is established.

In this thesis, we explore the challenge of giving a complete picture of people's online behavior across multiple OSNs by aggregating layers of online dimension and then investigate the effect of offline interactions on online link creation.

Collecting and aggregating available data about an individual from different platforms is a first step to capture an all-around picture of people's online behavior. We first perform a multidimensional analysis of users' behavior across OSNs on a dataset gathered from the social media aggregator Alternion. Then we focus on developing identification approach across OSNs to integrate information about an individual from different sources beyond social media aggregators.

Meetings and events are one of the favorite ways to get new friends in real life. But the mechanism and the extent of the impact of offline meeting events on the creation of online friendships have not yet been studied. We exploit these connections between offline and online dimensions to investigate the impact of offline sociality on the online.

As a matter of fact, we consider a small subset of the dimensions which may also have an impact on many aspects in the realm of computer science (e.g. recommendation systems, advertising, content dissemination, crowdsourcing, social discovery, etc.).

1.1 People's behavior across social media

Most of the studies on people's behavior in the OSNs assume that people get connected and establish relationships on a single social platform. This approach, however, enables the achievement of a partial representation of the online social behavior of individuals [59] because it is becoming evident that individuals are used to expressing their sociality through multiple layers, each associated to a specific medium, ranging from face-to-face and on-phone communications to a variety of online social networks.

To face the new challenge of giving an all-around picture of people's online behavior, we perform a multi-faceted analysis of users' behavior across multiple social media sites.

Contributions

Only few works [9, 47, 1, 45] have analyzed the users' behavior across online social networks, although they focus on specific behaviors as tagging or changes in clickstream. One of the main barrier in this field is the lack of datasets enabling the analysis of the single individual's behavior across different sites. Our study relies on a new rich dataset gathered from the social media aggregator Alternion. The novelty of Alternion dataset is its multidimensional and longitudinal nature. The dataset includes 19.680 distinct profiles from different countries and in different languages. The dataset contains, for each individual, a list of her/is favorite social sites and a multidimensional times series of posts, where each dimension corresponds to a specific social site. In addition, each profile reports the degree of the user on the social networks that make this information available. However, we are not able to get the node neighborhoods since Alternion does not return the neighbors' usernames. Alternion dataset enables us to answer our research questions by performing a variety of investigations that will mainly provide an in-depth understanding of the social behavior of the people.

The first research question regarding the multiple adoption of social sites is: How common is the use of multiple social media sites? We quantify this phenomena and compute how many different social websites people are able to manage. Membership distribution of users across sites has been studied in [71] only, where authors showed that it follows a power-law distribution from a dataset gathered in 2008. We extend that result, by confirming it on a more recent and larger dataset, which relies on the current social platforms. Our results show that on average a user is simultaneously registered on 5 social sites, while more than 95% of the users have joined 17 platforms at most. In fact, people are expressing their identity and their behaviors through multiple communication media. Are the popularity of these users equally distributed across the social sites in use?

Studying individual popularity across sites can be useful for designing the popularity prediction algorithms, which in turn can help sites determine users with the highest priority for friend recommendation algorithms. We verify if statistically significant correlations between the degrees of the same group of users in different sites exist. The presence of these correlations measures whether or not popular users in one site maintain their centrality across media. The way the users distribute their popularity across sites has been studied in [72] and [14]. However their analysis are different from the ones we have done on Alternion. Buccafurri et al. [14] using very limited dataset just report the average number of Twitter and Facebook friends. Zafarani and Liu [72] investigate how the maximum and minimum number of friends that individuals have across sites changes as users join sites. For the first time in the literature, our analysis focuses on finding correlations between the popularity of the same group of users in different sites. The analysis indicates that the user's popularity in a given social site barely corresponds or does not correspond at all to his/her popularity on another social platform. This means that users may have a very different centrality across the services, i.e. a single user might be a hub on one system and loose part of its hubbiness on the other. Does the same correlation also exist between the activity level of the same group of users across different platforms?

We measure the activity level of a user within a social media by means of his/her posting activity. While the research presented in [1] [47] investigates the individual tagging behavior in a small set of platforms, they mainly focus on studying the dependency of the intensity and the variety of user tagging activities on the features of the platform. To the best of our knowledge, there is no research analyzing the posting activity of users across different OSNs. Our posting activity analysis reveals that unlike the above discussion on centrality, there is a more evident positive correlation between the posting activities across social sites. The obtained values do not mean that users are equally active on social media, however there is slightly positive tendency to be active in different social sites. Do these results imply that users post on multiple social media every day?

We introduce the post multiplexity index to measures the propensity of a user of being multidimensional. The results show that most of the users tend to prefer a single media per day while sometimes they adopt multiple social media for posting. Is the activity dynamics of users uniformly distributed in time or does the burstiness characterize how the people post on online social media?

We analyze time-series associated to the posting activity of the single users across multiple social platforms and unveil their statistical properties focusing on measures which describe their level of burstiness. The results show that i) singularly each event sequence keeps an high level of burstiness, i.e. the bursty behavior of the aggregated interaction sequence is the union of bursty event sequences and; ii) a period of high activity in the aggregated sequence does not imply that the user is highly productive in each social media during the same period.

As far as we know, this study represents the first attempt to deal with the people's activities on multiple social media by using a large set of social platforms and users. The results represent novel insights about people's behavior across social media.

Chapter 2 is based on the following publications:

M. Zignani, A. Esfandyari, S. Gaito, G.P. Rossi, "Following people's behavior across social media", In proceedings of the 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pages 428-435, 2015.
M. Zignani, A. Esfandyari, S. Gaito, G.P. Rossi, "Walls-in-One: usage and temporal patterns in a social media aggregator". Applied Network Science, 2016 ; 1(1):1-24.

1.2 User identification across online social networks

To take advantage of the full range of services that online social networks (OSNs) offer, people commonly open several accounts on diverse OSNs where they leave lots of different types of profile information. The integration of these pieces of information from various sources can be achieved by identifying individuals across social networks. The ability to gather the public traces left during the online activities would lead to a deeper understanding of the user's identity and behavior. This would improve service provisioning, enable service customization and cross-domain recommendation, and give rise to other services in different domains.

In chapter 3, we address the problem of user identification by treating it as a classification task.

Contributions
First, we rely on common public attributes that are available through official APIs of social networks to overcome the privacy issues. Second, we show that using standard approach in the literature for selecting negative instances results in a high number of false positives in practice. Thus, we propose different methods for building negative instances going beyond usual random selection in order to investigate the effectiveness of each method in training the classifier. Third, two test sets with different levels of discrimination are setup to evaluate the robustness and accuracy of our different classifiers. Finally, we measure the effectiveness of the approach in real conditions by matching a profile dataset extracted from Google+ with two other datasets obtained from Facebook and Twitter.

Chapter 3 is based on the following publication:
A. Esfandyari, M. Zignani, S. Gaito, G.P. Rossi," User Identification across

Online Social Networks in Practice: Pitfalls and Solutions". Journal of Information Science (October 2016, DOI: 10.1177/0165551516673480)

1.3 The effect of offline sociality on online interaction

As link creation in online social networks is one of the key-points to understand the network growth, its mechanisms have been largely investigated. Studies well mainly devoted to discover the mechanisms built on features intrinsic to the network itself, as common neighbors. The role played by users' offline meetings on their online friendship creation has often been argued[56], but it has not yet been studied, mainly because of the lack of available dataset.

More and more often offline events are advertised on online social networks, bridging the offline to the online users' social activity. We leverage this connection between offline and online sociality to build a novel dataset of events advertised on Facebook which enables us to take a first step towards understanding the effect of offline events on online link creation by performing a temporal analysis on the social network built by people attending the event. In particular, we aim at answering the following questions: How can participating in an event in Facebook, that is usually followed by physical attending on the day of the event, change the structure of the online social network? How do friendship relations among the attending people increase during the event?

Contributions

In chapter 4, we perform a temporal analysis of the *event social network*, constituted by people declaring to attend the event advertised on the Facebook and the links between them. We explore how the network evolves during the event time period. To evaluate the impact of an event on the social network of its attendees, we investigate how the networks change at macroscopic level by means of the network communicability, at mesoscopic level by analyzing the clustering coefficient trend and at microscopic level by observing the increase of users' degree. We analyze the results and reveal the effect of the events on the new friendship creation between the attending people.

Chapter 4 is based on the following publication:
A. Esfandyari, M. Zignani, S. Gaito, G.P. Rossi, "Impact of offline events on

online link creation: a case study on events advertised on Facebook", Proceedings of the 31st ACM/SIGAPP Symposium on Applied Computing (SAC), 2016.

Chapter 2

People's behavior across social media

2.1 Introduction

During the last decade, large datasets describing online social networks have been made available together with an extensive literature which enabled a comprehensive description of OSNs. Most of these studies assume that people connect and establish relationships on a single social platform. This approach, however, enables the achievement of a partial representation of the online social behavior of individuals [59]. It is becoming evident that individuals are used to expressing their sociality through multiple layers ranging from face-to-face and on-phone communications to a variety of online social networks. Especially in their online life, users have access to a wide portfolio of social platforms which allow them to differentiate their interests and convey diverse contents. As a result, a user can be registered to many social media and interact with her/his friends' circles throughout different channels. For instance, a user can share her/his photos on the Instagram, organize an event on Facebook and maintain the relationships with workmates on LinkedIn. These arguments are quantitatively reported in a survey conducted in September 2014 [1] where it has been highlighted that the adoption of multiple social platforms is on the rise. In fact, 52% of online users now use at least two social media, representing a significant increase since 2013, when the figures were close to 42%. The diversification in the usage of the social platforms, in conjunction with their specialization, is causing a fragmentation of users' identities among diverse social media. This process of identity fragmentation makes the understanding

[1] http://www.pewinternet.org/2015/01/09/social-media-update-2014/

of the online behaviors more difficult since the data are split and need to be matched and fused.

In this chapter, we move towards the reduction of the above segmentation by facing with the new challenge of giving a complete and compelling picture of people's online behavior. To this end, we move within the multidimensional or multiplex network theory [44, 12, 10, 34, 17] since it provides the tools and the models which better capture and measure the interplay/correlation among the social sites' users adopt. Specifically, the scenario we are dealing with is well represented by a multigraph [74], a special case of an heterogeneous information network [62]. So far the multiplex network theory has been applied to different real case studies, from citation [40], co-author [61] and conference-author networks [63] to power grids [13], economic [41, 42, 6] and biological networks [8]. However most of these studies assume that the underlying multiplex network is static or well-known dynamical processes occur onto them. On the contrary, little is known about the temporal interactions of people when they have different communication media available, especially in the online world. So the main goals of this research are *i*) applying the multiplex network theory to understand the on-going phenomenon of the adoption of multiple social site; and *ii*) measuring how this process impacts on the interaction dynamics.

In practice, the study of the behaviors of people across different online social networks is at its very beginning. In fact, while many works have been published about profile matching algorithms across sites [70, 15, 66, 23], only a few works [9, 47, 1, 45] have analyzed the user's behavior across sites, but they focus on specific behaviors as tagging or changes in clickstream.

One of the main barriers in this field is the lack of datasets enabling the analysis of the single individual's behavior across different sites. This criticality can be ascribed to a variety of motivations. First of all, people are unwilling to make all information about their online social life public and, secondly, social site providers have less of an interest in providing tools to integrate other social platforms. To overcome the data problem, in this study we rely on the services offered by social media aggregators. A social media aggregator is a Web service which allows users to collect and manage different social site accounts through a single application. Most of these services offer to their users a private space, where they can share a content simultaneously on multiple media. Among these services, we retrieve data about users and their social sites from Alternion[2], because it allows the users to make information about their social sites public, unlike most of social media aggregators. Thus, it is possible to collect data about their profiles with personal public information and the posted contents.

This research offers a first data-driven contribution to the study of people's behaviors across social sites while addressing the following research questions:

- **Q1**: How common is the use of multiple social media sites? Does the same hold for active users?
- **Q2**: Is a person's popularity uniform, *i.e.* more or less equally distributed across the social sites in use?
- **Q3**: Is users' production alike in different social media ? Do the most active users behave similarly on different media?
- **Q4**: How do people manage their posting activities during the day? Do they post on multiple social media every day or do they alternate in the choice of the publishing platforms?
- **Q5**: Are users coherent in the choice of their usernames, keeping their identifiability across social sites?

By answering the above questions, we introduce the following main findings and contributions. On the one hand, some findings confirm that the adoption and the active usage of multiple social sites is gaining momentum and maybe be studied through social media aggregators. On the other, we find that a full multiplexity of the interactions and posting activity is difficult to reach in short periods (day) but more evident in longer.

- **Membership distribution across social media.** The way users distribute their membership across sites has been recently studied in [71] and has been the subject of a few market surveys. We confirm and extend these results by means of a more recent dataset, which consequently relies on the today's most popular social websites. The usage of multiple social platforms, raising in 2008, is now strengthening and social media aggregators offer a chance to collect data about this phenomenon.
- **Popularity across social media.** By performing a correlation analysis we investigate the maintenance of users' popularity across social sites. The analysis led to not straightforward results and this indicates that user's popularity in a given social site barely corresponds or does not

[2] http://www.alternion.com

correspond at all to his/her popularity on another social platform. The lack of a strong degree correlation has been observed in different contexts such as online/offline interactions [20], virtual worlds [64] and other multiplex networks [52]. Here, we observe the same phenomenon in online social platforms.

- **Posting activity across social media.** We aim at understanding whether an active and productive user on a social media preserves his/her aptitude on the other social media. Correlations between the posting rates on different social sites show that an easy answer is not possible, although we measure slightly positive correlations for many couples of social platforms. These results represent one of the main novel insights of this research since little is known about the users' engagement in multiple social media.

In general, the last two findings stress the fact that people, in a multiplex scenario, change their importance and preferences from medium to medium. The ability of handling some particular media better than others, the different interests and the usability of the services may be the reasons behind this behavior.

- **Burstiness in temporal patterns.** We have analyzed the posting inter-event times both aggregated on all social sites, to get an overall picture of the activity dynamics of users, and a per site evaluation, to get temporal patterns specific to a social website. This way, we evaluate the dynamics of the online activity by a true multidimensional approach. We discovered that the posting activity on online social media is bursty and highly heterogeneous. In particular, the bursty behavior of the aggregated sequence is the union of bursty posting event sequences on different social media. Moreover, a period of high activity in the aggregated sequence does not imply that the user is highly productive in each social media along the same period.
- **Temporal multiplexity.** We introduce the *post multiplexity* index to measure the propensity of a user to being multidimensional. The results show that most of the users tend to prefer a single media per day while sometimes they adopt multiple social media for posting.

The above results represent the first steps towards a multidimensional approach in the study of human dynamics [65] in communication and social networks. Some attempts to extend and model bursty dynamics can be found in [57] and [31], however the layers are few or do not really represent communication channels. In this chapter, thanks to a large set of social media, we highlight the multidimensional bursty nature of human dynamics in multiple social sites. Specifically, we show that the overall burstiness is a consequence of a complex mixture of non-stationary interests to the chosen social media.

- **Patterns in username usage.** We evaluate how users are coherent in the choice of their username across social sites. Results show that users often maintain the same username across different social websites, but are more likely to change them among websites whose norms and scopes are different.
- **Alternion datasets.** The above results rely on two new collected datasets which capture multidimensional and longitudinal information about how online users behave across multiple social platforms. To the best of our knowledge, they represent the most updated available datasets which combine posts, their contents and the multiple profiles of a large set of people. The Alternion dataset allows us to quantify the multiple platforms usage without requiring a periodical survey; on the other, additional information about posts could be extracted to verify whether the same observations hold for users' engagement. Furthermore, posts provide the temporal information necessary to the study of multidimensional human dynamics in online social networks.

The aforementioned results support the increasing awareness that single social site studies provide a very partial description of human social behavior which effectively needs a multisite approach to be described and fully understood. To this end, social media aggregators, such as Alternion or About.me, represent data collection to be deeply analyzed, provided that we take into account the bias given by the typical users of these platforms.

2.2 The related works

Several studies have been performed to compare the users' behavior among different OSNs. Geo et al.[22] compare users' behavior on two different microblogging platforms, Sina Weibo and Twitter. They analyze the access behavior and the textual features of microposts. Moreover, they investigate the temporal dynamics of microbloging behavior including the shift of user interest over time.

Gyarmati and Trinh [26] analyze the characteristics of the activity of the users of Bebo, MySpace, Netlog, and Tagged. They provide statistics of the user's behavior on a daily timescale.The main findings of the article conclude that users' online time spending can be modeled with Weibull distributions; and the duration of OSN users' online sessions shows power law distribution characteristics.

Anh et al.[2] analyze sample networks from Cyworld, orkut, and MySpace in terms of degree distribution, clustering coefficient, degree correlation, and average path length.

Zhao et al.[73] compare the content of Twitter with a traditional news medium, New York Times, using unsupervised topic modeling. They find that although Twitter users show relatively low interests in world news, they actively help spread news of important world events.

Dwyer et al.[18] compare the attitudes and behavior of the users between Facebook and MySpace. They show that the Facebook members are more trusting of the site and its members, and more willing to include identifying information in their profiles. Yet MySpace members are more active in the development of new relationships.

Hughes et al.[28] examine the personality correlates of social and informational use of Facebook and Twitter. They show that personality is related to online socialising and information seeking/exchange. Their results also reveal that a preference for Facebook or Twitter is associated with differences in personality.

Mislove et al.[49] present a large-scale measurement study and analysis of the structure of four online social networks: Flickr, YouTube, LiveJournal, and Orkut. Their results show that social networks are structurally different from previously studied networks and have a much higher fraction of symmetric links and also exhibit much higher levels of local clustering.

All of above mentioned studies do not consider the behavior of the same group of people in different Systems. In fact, while many works have been published about profile matching algorithms across sites [70], [15], [23], [66], [30], only a few works [9], [47], [1], [72], [14] have analyzed the users' behaviors across sites.

Benevenuto et al. [9] have performed a study on a clickstream dataset collected from a social network aggregator, providing users with a single interface for accessing multiple social networks. They analyze how a specific behavior changes across sites without considering users that are shared across sites. While Alternion includes information of the same group of people registered in different social networks, including times series of posting activity in each OSNs, centrality in each OSNs and other profile information in each OSNs, Benvenuto et al.[9] dataset includes only clickstream information of different users on different social network.

Meo et al. [47] collect datasets of tagging statistics of 1,467 users have profiles on Flickr and Delicious and only 321 users have profiles in all the three systems. While their dataset includes the number and the content of tags for users in each platform, the time series of tags are not available. In compared to this dataset, Alternion is a much more complete one containing the profile information of 19.680 that enables analyzing various aspect of users behavior across OSNs. Specifically, the number of social platforms amounts to 152, from the most famous Facebook, Twitter to the less common Discus or Zazzle. They study the characteristics of the user's profiles from Flickr, Delicious and StumbleUpon with respect to three different aspects: (1) the intensity of user tagging activities, (2) tag-based characteristics of the user's profiles and (3) the semantic characteristics of the user's profiles.

Abel et al. [1] focus on presenting the system Mypes. Mypes supports the linkage and aggregation of user profiles available in various Social Web systems, such as Flickr, Delicious and Facebook. Their dataset consists of 3080, 3606, 1538, 2490 and 15947 public profiles from Facebook, LinkedIn, Twitter, Flickr, and Google, respectively. They crawled profiles attributes such as first name, last name and email, from each service. While Abel et al. [1] investigate the completeness of individual and aggregated Profiles attributes, we focus mainly on the membership distribution, popularity and posting activity analysis on our dataset Alternion. Among the users for whom they crawled the Facebook, LinkedIn, Twitter, Flickr, and Google profiles were 338 users who had accounts on all of these five different services. They use this part of dataset to analyze the individual's tagging behavior in different systems.

Zafarani and Liu [72] study the friendship and popularity of the users across social media sites on a dataset of 96,194 users, each having accounts on different sites. They show that the maximum number of the friends individuals have across sites increases linearly as the users join sites and their minimum drops exponentially. Their analysis also reveals that the users' joining multiple sites cannot increase their average popularity and that the average popularity converges to a fixed value as users join sites. For each individual in the dataset, they have the number of friends a user has across different sites. Alternion dataset includes more information rather than the number of friends.

Buccafurri et al. [14] study Friend distribution and the attitude of users to have overlapping friendship relations on a dataset of 757 users, each having accounts in both Facebook and Twitter.The dataset includes the number of the friends that users have in both Twitter and Facebook. Moreover, it contains a binary value for each user indicating whether the profile of the user in Facebook is public or not. For privacy analysis, they count how many users of the sample with two accounts choose to disclose their Facebook information on the social network, thus making their Facebook account public.

Although, the recent studies that have analyzed the users' behaviors across OSNs are more relevant to this work, they consider limited aspects of human behavior. All of the datasets in the literature are differ from our dataset (Alternion) and cannot answer our research questions. Our analysis is based on datasets of the same group of users with profiles on different social media to have meaningful and well-founded results. We investigate those aspects of human behavior across online social networks that have not been studied in the literature.

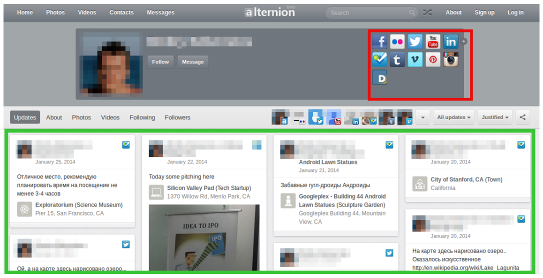
2.3 Methodology

Nowadays, people have at their disposal a wide selection of online social sites, each having its own peculiarity. This way, the adoption of multiple social platforms by the same person is becoming increasingly spread. Now, people can exploit and combine their favorite social media according to their needs. For example, a tourist could share his/her position by Foursquare, meanwhile s/he uses Twitter to communicate his/her mood and shares on Instagram a selfie with the "Gioconda".

At the same time, people have recently shown a growing interest in tools managing their online life in a centralized way, *i.e.* social media aggregators. Social media aggregators allow the users to merge their own identities into a single profile by gathering their online activities from different social platforms, such as Twitter, YouTube, LinkedIn, Facebook, and many others. The aggregation is enabled by APIs provided by social networks, so that people may have control on the data the aggregation platform can access.

Most of these services offer a private space to the users, where they can share a content simultaneously on multiple media, but profiles with public contents and public API are not available. The media aggregator Alternion [3] stands out because, unlike most of similar services, allows users to decide which information about their social sites - including profile information and contents - can be made public. Furthermore, the service retrieves data from more than 200 social sites and manages social relationships among Alternion profiles.

2.3.1 Data collection methodology



January 25, 2014

Отличное место, рекомендую планировать время на посещение не менее 3-4 часов

Exploratorium (Science Museum)
Pier 15, San Francisco, CA

Ой, а на карте здесь нарисовано озеро...

January 22, 2014

Today some pitching here

Silicon Valley Pad (Tech Startup)
1370 Willow Rd, Menlo Park, CA

IDEA TO IPO

IPO

Android Lawn Statues
January 21, 2014

Забавные гугл-дроиды Андроиды

Googleplex - Building 44 Android Lawn Statues (Sculpture Garden)
Googleplex Building 44, Mountain View, CA

January 20, 2014

City of Stanford, CA (Town)
California

January 20, 2014

На карте здесь нарисовано озеро..
Оказалось искусственное
http://en.wikipedia.org/wiki/Lake_Lagunita

Fig. 2.1. **An example of a profile page in Alternion.** The red box highlights the social platforms shared by the user. A pop-menu shows the number of relationships in each social network. The green box highlights the area containing all the public posts gathered by the API of the services.

Since the service does not expose a public API, we developed a crawler to retrieve the Alternion profiles and their public updates. An example of an Alternion profile page is shown in Figure 2.1. The page exposes the social sites associated to the Alternion identity. Each icon links to the relative profile and shows the username chosen in the target social site. The "Updates" tab

[3] http://www.alternion.com

reports the public posts grouped by social sites. The content of each post highly depends on the information released by APIs.

We collected registered users by exploiting a search function which returns a different set of 40 random users every two minutes. Data collection started on October 2014 and resulted in 19.680 distinct profiles from different countries and in different languages. From a user's profile, we extract the set of social sites a user associates to the profile and all the public contents s/he had published since the registration date. For each content, we gather information about the social media used to post it, the content itself according to the format returned by the APIs and the publication date. Furthermore, not to overload the service and to respect politeness, we limit the number of contents for each user to 10.000 at most. In practice, the dataset contains, for each individual, a list of her/is favorite social sites and a multidimensional time series of posts, where each dimension corresponds to a specific social site. All the analysis, but username patterns reported in the last section, will be performed on this dataset $D1$. In addition, each profile reports the degree of the user on the social networks that make this information available. However, we are not able to get the node neighborhoods since Alternion does not return the neighbors' usernames. Consequently, we are not able to build any network from the information provided by the social media aggregator. Since we may be able to build different types of relation by exploiting implicit links such as mention, we model our dataset by a directed multigraph $\mathcal{D} = (V, E, D)$. Each user (identity) is an element of V, while $E \subseteq V \times V \times D$, with D the set of social media in Alternion, is the set of directed multi-edges. Finally, we associate to each user $u \in V$ a sequence of timestamped events (t, pe_d), where pe represents a post published on the social media $d \in D$.

For the study of username patterns only, we collected a different dataset, $D2$, whose characteristics are specifically designed to this aim. We adopted a more targeted sampling approach by collecting 15.000 profiles with an English first name [4] to make the text analysis of usernames more affordable. This way we maintain the alphabet accordance between the username (typically ASCII character) and the information about the identity (first and last name are usually written in mother tongue). Finally, the dataset $D2$ contains for each individual a list of her/is usernames.

[4] http://census.gov

2.3.2 Dataset characteristics and representativity

The dataset $D1$[5] contains the profile information of 19680 users, each reporting the platform Alternion among her/is social sites. Specifically, the number of social platforms amounts to 152, from the most famous Facebook, Twitter to the less common Discus or Zazzle. In Figure 2.2, we report the number of users who share their profile information of a social site, while in Figure 2.3 we report the number of users for the top ten social platforms. As expected, and ignoring Alternion, Facebook is the most used social site and the top 10 correspond to the most popular social media.

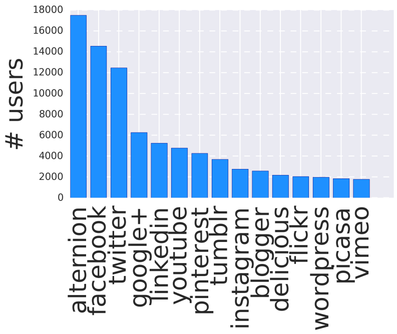


Fig. 2.2. **Social site adoption.** The number of users for each social platform.

In Figure 2.3, we also report the number of active users per social site. An active user refers to a person who has published at least one content on the profile page. About half of the profiles (9829) are active and do prefer Twitter as publishing media. Although the Twitter result is expected due to the Twitter intrinsic nature (interest and media broadcasting network), LinkedIn and Google+ are more favorite than Facebook when we consider active users, even if the continuous releases of new versions of the Facebook API may influence the available information. Anywise we generally note that the users who actively adopt social platforms always represent a fraction of the registered users.

We obtain a deeper understanding if we combine the above results with those presented in Figure 2.4, where we report the number of posts grouped by

[5] The dataset is available by e-mail.

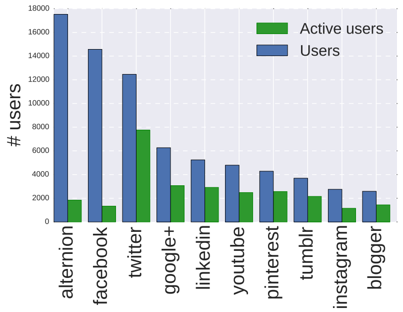


Fig. 2.3. **Social site engagement.** The number of users/ active users for the top ten social sites. Active users represent people who have published at least one post on a social site.

social sites. Here, we note that the Facebook active users are more productive than Google+ users. In fact, although they are fewer, Facebook users publish more or less the same amount of posts as Google+'s. Finally, the results in Figure 2.4 confirm the predominant role of Twitter in the production of contents and posts, *i.e.* more than 6 million posts over an overall amount of more than 8.5 million posting events.

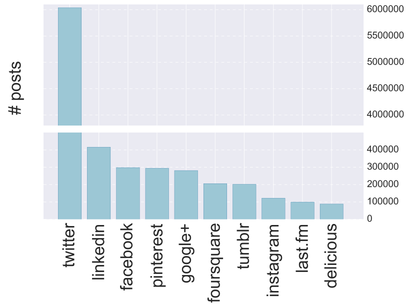


Fig. 2.4. **Posting activity.** The number of posts for each social site.

As regards the posting activities of the active users, in Figure 2.5, we show the distribution of the number of published contents per user. The post sampling covers the time interval from 10 July 2005 to 10 November 2014. The distribution seems to obey to a heavy tail with exponential cut-off. During the examined period, we observed that the users have been quite productive:

on average, the users have published about 800 different contents and half of the users have produced more than 300 updates.

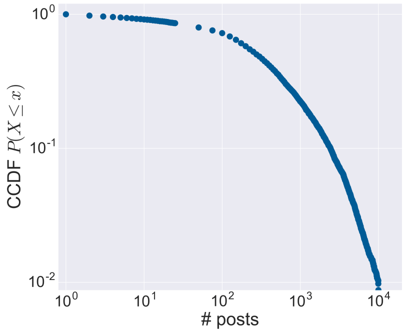


Fig. 2.5. **Posting activity.** The complementary cumulative distribution function (CCDF) of the number of posts published by each user.

The above results highlight that Alternion effectively captures a representative sample of today's users of social media sites, in terms of favorite social sites and publishing media as compared with public available data [1] on social media usage. Indeed, the usage proportion across the social sites is almost equal to those collected by aggregating the official statistics released by the different social platforms. Despite data show an alignment between the Alternion users and the users of the different social media, the representativeness issue of the dataset still persists. For instance, the users' demography is unbalanced towards western and English-speaking countries, the contents posted on the Alternion profile depend on the API provided by the social media. Lastly, the adoption of social media aggregators is more common and useful to an audience active on social media. Nevertheless, it represents the most recent dataset able to capture how users behave across multiple sites and an easy tool for studying the on-going phenomenon of the multiple adoption of social platforms.

2.4 The Usage of multiple social sites

In a survey conducted in September 2014 [1], it has been highlighted that the adoption of multiple social platforms is on the rise. In fact, 52% of online users now use at least two social sites representing a significant increase since 2013, when the figures were close to 42%. On the one hand, the Alternion dataset allows us to quantify this phenomenon without requiring a periodical survey; on the other hand, additional information about posts could be extracted to verify whether the same observations hold true for the users' engagement. Indeed, being registered to a service does not always imply using it.

To answer Q1 - *How common is the use of multiple social sites? Does the same hold true for active users?* - we analyze how the users are distributed across sites. The main goal of a social media aggregator is to group different social sites into a single access point. This characteristic allows us to compute how many different social websites people are able to manage. How users distribute their membership across sites has been studied in [71] only, where the authors have shown that it follows a power-law distribution from a dataset gathered in 2008. Our goal is to extend that result, by confirming it on a more recent and larger dataset, which relies on the current social platforms. In Figure 2.6, we report the probability distribution function (PDF) of the number of sites joined by a given person. On average, a user is simultaneously registered on 5 social sites, while more than 95% of users have joined 17 platforms at most. Alternion results are consistent with the aforementioned survey; indeed about 56% of Alternion users use at least three different social platforms[6]. The behavior of active users gives a more direct and clear evidence

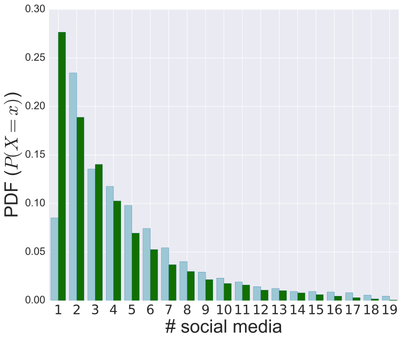


Fig. 2.6. **Social sites per user.** The PDF of the number of social sites joined by each user (blue)/active user (green).

[6] Alternion account is included by default, so counting starts from 1

of the adoption of multiple sites. In fact, the publication of contents on a specific social media confirms its adoption in practice and it is strictly related to the user's engagement. This way, we compute the distribution of the number of social platforms being simultaneously used by each active user. As shown in Figure 2.6, the observations about the multiple usage gets stronger: 73% of the active users publish on at least two social networks. In general, the propensity to adopt multiple social sites is becoming increasingly ingrained amongst online users, especially if we consider the effective publishing on different platforms.

The above analysis on the people's propensity to be active on multiple social media does not take into account the amount of the contents published on each platform. To this end, we analyze the productivity of the users who join more than one site simultaneously. Let $p_1, p_2, \ldots, p_n$ indicate the number of the posts of a user on the n sites where s/he is active and $\bar{p} = \sum p_i/n$ the average number of posts per site. For each active user, we compute $\bar{p}$ and group them by the number of multiple sites they have joined; then we calculate the mean and the standard deviation of $\bar{p}$ for each group. In Figure 2.7, we report the above quantities as a function of the number of sites. The figure indicates that *i*) the more sites a user joins, less posts for the site on average s/he publishes, and *ii*) the users who post on few sites are more heterogeneous in their activity since the standard deviation decreases as the number of the sites increases. We suppose that the first remark is a consequence of the limited human resources in terms of usable time and ability of handling multiple tasks at the same time [48].

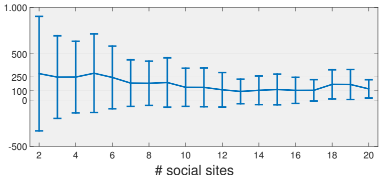


Fig. 2.7. **Posts and number of sites.** The average $\bar{p}$ as a function of the number of sites users have joined. The error bars report the standard deviation.

Finally, in Figure 2.8, we show the number of the users who have joined different social sites by considering two classes: social media and social networks. The former denotes the social sites where relationships are implicit, the latter indicates the social sites where the relationships are explicitly defined. These figures show that more than 90% of the users have registered to at most 5 sites and some of them share information on more than 10 social media. In general, we observe that it is more likely to adopt social media than social networks.

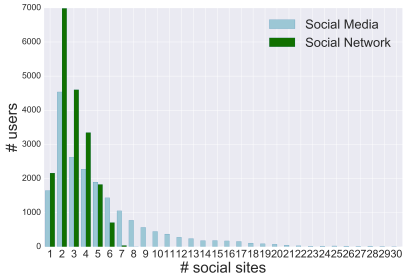


Fig. 2.8. **Social media and networks.** The PDF of the number of social sites joined by each user subdivided in social media and social networks.

2.5 Centrality and activity correlation across social sites

To answer Q2 - *Is a person's centrality uniform,* i.e. *more or less equally distributed across the social sites in use?* - we treat the user degree on different social sites as an index of centrality in the network. This information, reported on the Alternion profile, is not available for every site since just a few APIs allow to retrieve this information. So, we limit our analysis on the degree to Facebook, Twitter (in/out), LinkedIn and YouTube (in/out).

Q2 is meant to verify if statistically significant correlations between the degrees of the same group of users in different sites exist. The presence of these correlations measures whether or not popular users in one site maintain their centrality across media. If we denote k_u^s as the degree of the node u in the site s, we can evaluate the degree of correlation between pairwise social sites by adopting different methods [11]. First we compute, for each pair of social sites, the joint distribution $P(k^{s1}, k^{s2})$ to obtain a characterization of the relations between the degree sequences. For example, in Figure 2.9, where we show the joint distribution for Facebook and LinkedIn, we observe a slightly positive correlation, especially in the left bottom part of the distribution (dark blue regions). In particular people with about 200 friends in Facebook and about 20/25 relationships in LinkedIn are more likely than others.

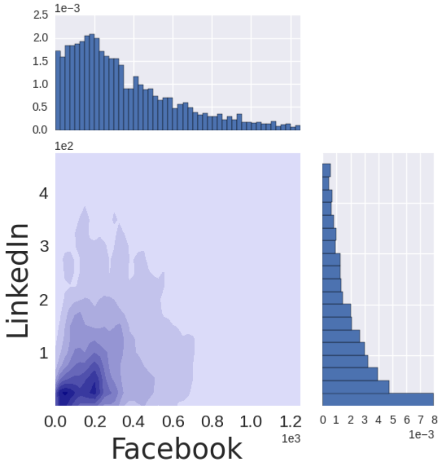


Fig. 2.9. **Facebook and LinkedIn.** The joint probability distribution estimated by kernel density method.

By analyzing the joint distributions only, it is difficult to compare the relations between a social site against the other ones. For this reason, we compute the average degree of a node in a social site s_1 conditioned on the degree in the site s_2. The resulting computation for LinkedIn conditioned on Facebook has been shown in Figure 2.10. In this case, we observe an initial increase of the average degree in LinkedIn as a function of the Facebook degree up to about 500 friends. Then we observe a more unstable trend also due to a low number of data points. In the Figure 2.11, we report a more pronounced lack of correlation involving the out-degree in Twitter and YouTube. Here, we cannot establish an increasing relation between the out-degrees, so people how follow many users in Twitter would not follow the same amount of channels in YouTube. The specificity of the two social platforms, *i.e.* news broadcasting and video sharing, could be the cause for the weak correlation.

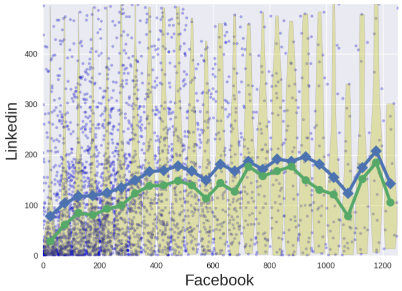


Fig. 2.10. **Facebook and LinkedIn.** The conditional distributions of the LinkedIn degree given Facebook degree. For each 50 size bin, we show the degree distribution through the violin plot with its average (blue) and its median (green).

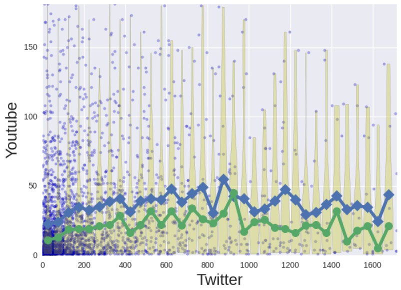


Fig. 2.11. **Twitter and YouTube.** The conditional distribution of the out-degree in YouTube given the Twitter out-degree. For each 50 size bin, we show the degree distribution through the violin plot with its average (blue) and its median (green).

Keeping Facebook as conditioning social site, we observe that the average degree increases in some media like LinkedIn and Twitter, while the in/out degrees in YouTube are uncorrelated with the Facebook degree; as shown in Figure 2.12. The comparison between the degree and the in/out degree is dictated by the inability to extract mutual links (more similar to a friendship link in Facebook or LinkedIn) since we can only retrieve the counting of the followers/followees in Twitter and YouTube. Finally, to get an overall picture of the pairwise degree correlations, we apply a rank correlation analysis on the different pairwise sequences. Rank correlation analysis allows us to test if the ranking induced by the different degrees is similar or not. As a rank

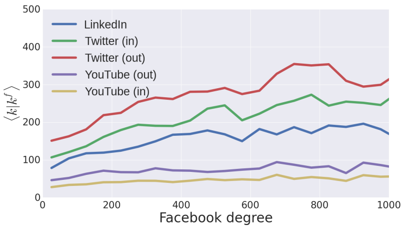


Fig. 2.12. **Degree dependencies.** The average degree in LinkedIn, Twitter and YouTube as a function of the degree in Facebook.

correlation method, we compute the Kendall's rank correlation coefficient τ_b[7] on the ranking induced by the degrees. In Figure 2.13, we visualize the rank correlation matrix, where each row (column) corresponds to a different social site. A strong positive correlation does not exist, rather the scenario is multifaceted. In most pairs, there is only a limited positive correlation $(0.1 - 0.23)$ between degree centralities. This means that users may have a very different centrality across the services, *i.e.* a single user might be an hub on one system and loose part of its hubbiness on the other. One reason may rely on the different goals of the services; whereas LinkedIn is business-oriented or Twitter is an interest network, Facebook incorporates all the previous features. So, for instance, LinkedIn may only capture a part of the Facebook friends.

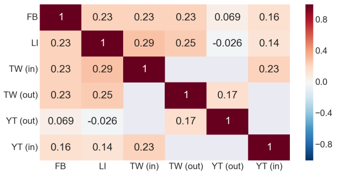


Fig. 2.13. **Degree rank correlation.** The correlation matrix among the degree sequences in the different social sites.

The above conjecture cannot be verified since we are not able to compute the intersection between a node's neighborhoods in different social networks (information not provided by Alternion). However we can quantify and analyze the difference between the neighborhoods in terms of their size. To this end,

[7] τ_b takes into account ties.

given a user u and two sites s_1 and s_2, we define the *friend deviation* Δk_u^{s1s2} as:

$$\Delta k_u^{s1s2} = k_u^{s1} - k_u^{s2} \tag{2.1}$$

We compute the friend deviation between Facebook and Linkedin(Δk_u^{FL}), Facebook and Twitter(Δk_u^{FT}), Facebook and YouTube(Δk_u^{FY}), Twitter and LinkedIn(Δk_u^{TL}) and Twitter and YouTube(Δk_u^{TY}) for the users who have joined them. We report the trends of the friend deviations in Figure 2.14 sorted in decreasing order. We observe that Δk_u^{FT}, Δk_u^{FL}, Δk_u^{FY}, Δk_u^{TL} and Δk_u^{TY} are positive for 5360 users (out of 8527), 3313 users (out of 4361), 2944 users (out of 3177), 2518 users (out of 4148) and 2760 users (out of 3098), respectively. These results indicate that the users in our dataset prefer to create friendships in Facebook rather than in LinkedIn, Twitter and YouTube. Moreover, users prefer Twitter rather than LinkedIn and YouTube to establish friendships. One remarkable result is that the users preferring Twitter rather than Facebook have significantly more friends than those they have in Facebook. In general, maintaining the importance across social media is not a straightforward task and asks for a deeper understanding; for example it is not clear how user's neighborhoods in different media overlap.

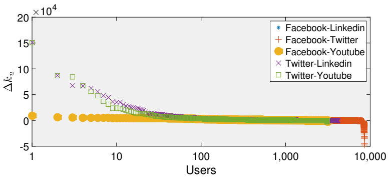


Fig. 2.14. **Friend deviation.** The friend deviation scores sorted by decreasing order.

Finally, we apply the above methodology *i*) to investigate how often the users publish contents in different social sites, and *ii*) to asses if a form of correlation exists between the amount of posts and the number of friends. For point *i*), we are verifying whether users who post a lot and often on a social platform, are equally active in other platforms (see Q3). We measure the activity level of a user within a social media by means of the posting rate, measured in number of posts per week. By considering the posting rate rather than the post count, we mitigate the effects given by the adoption of social media in different periods. We report the results about the analysis of the rank correlation matrix applied to pairwise posting rate sequences. In Figure 2.15, we visualize the Kendall's coefficient τ_b for the most used pairs of sites. Unlike the above discussion on centrality, there is a more evident positive correlation between the posting activity across social sites. The obtained values do not mean that users are equally active on both social media, however there is a positive tendency to be active in different social sites. In general, the maintenance of the posting activity across social media is not a straightforward task, like in the degree analysis. By the second point, we wonder if

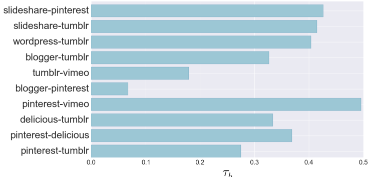


Fig. 2.15. **Post rank correlation.** Kendall's coefficients on the posting rate computed on some pairs of social media.

users with many friends in a OSN are more active and productive than people with fewer friends. For this reason, we consider only the users whose degree and posting rate are available and combine these information. The analysis of the Kendall's coefficient τ_b highlights a medium positive correlation between the two variables for all the online social networks ($\tau_b \in [0.27, 0.4]$) except YouTube ($\tau_b = 0.1$). High degree people tend to post and publish more than those with few friends.

2.5.1 A case study on 4 social media.

Up to now the analysis has focused on the pairs of social media. In this section we present a particular case study that involves 524 users who have joined Google+, Pinterest, LinkedIn and Twitter. We select this subset of social media because they have the highest number of users w.r.t. to other subsets of 4 elements and their users are also very active. For instance in Twitter users published 1025 elements on average, followed by LinkedIn (139.32), Google+ (137.55) and Pinterest (134.71). In fact, we observe that less than 40% of users in Google+, Pinterest and LinkedIn have more than 100 posts. While in Twitter almost 87 percent of the users produced more than 100 posts.

In the light of the results about the moderate correlation among the posting rates, we wonder if, in this group, people who actively post in a service necessarily produce many posts in the other services. To this end, we denote the top 5% of users in each media as most active users. Only two users are the most active in all the services. 36%, 30% and 42% of the most active users in Twitter are also the most active in Google+, Pinterest and LinkedIn respectively. 23% and 14% of the most active users in Google+ are also the most active in Pinterest and LinkedIn respectively. The above observations support the presence of users who are very active pairwise but, whereas the number of sites actively used increases, the number of posts and consequently the productivity reduces as observed in the previous section.

We quantify the diversity of the posting activity across social media by defining the posting deviation Δp_u^{s1s2} of a user u as:

$$\Delta p_u^{s1s2} = \frac{|np_u^{s1} - np_u^{s1}|}{max(np_u^{s1}, np_u^{s1})} \tag{2.2}$$

where np_u^{s1} (np_u^{s2}) denotes the number of posts of user u on the social media $s1$ $(s2)$. Δp_u^{s1s2} ranges from 0 to 1: if Δp decreases towards 0, u tends to publish the same number of posts both in $s1$ and $s2$. As shown in Figure 2.16, where we report the posting deviation Δp between Google+ and the other social networks, this quantity decreases linearly. From the figure emerge that users' behavior is more variable between Google+ and Twitter than the other social networks. Only 9% of users show a post deviation less than 0.5. Most of the users prefer to post on a single service. So, users who laboriously publish posts in one service do not publish with the same rate in the other ones.

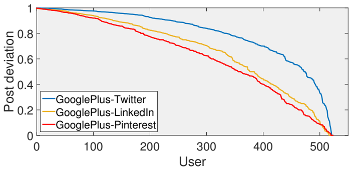


Fig. 2.16. **Post deviation.** The deviation of the number of posts.

2.6 Temporal patterns in the posting activity

It is a well established result that the human dynamics on communication media are bursty and scarcely uniformly distributed in time, even removing the effect of seasonal or circadian cycles [33, 57]. So we investigate whether the burstiness characterizes how people post on online social media, too. Figure 2.17a) displays a typical posting activity of a user adopting different social media simultaneously. The overall activity presents periods where the user is almost inactive, interleaved with periods of high activity. We analyze the statistical properties of the inter-event time Δt, i.e. the period between two consecutive posts published by the same users. Many studies have shown that the inter-event time distribution in different human processes [21, 5, 33] (e.g. mobile phone call, email, direct physical contact, link formation in online social networks) is heavy-tailed. In Figure 2.18 we report the distribution of the inter-event time resulting from the aggregation of the behavior of all the users. The distribution seems to obey to a heavy-tailed law and supports the hypothesis that the posting activity on online social media is bursty and highly heterogeneous.

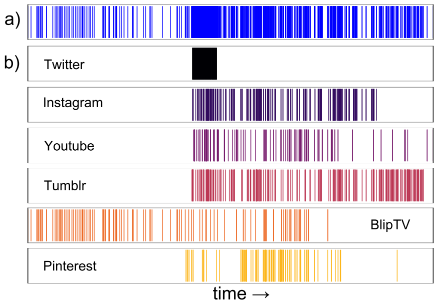


Fig. 2.17. **Posting activities on different media.** Posting activity on different social media. In a) the raster plot of the overall activity of a person who uses many social media simultaneously. In b) the user's activity split into the six social media s/he is using.

The opportunity of tracking multiple online activities simultaneously allows us to highlight the complexity of the users' posting activity, which appears like a compound combination of the single activities on the different

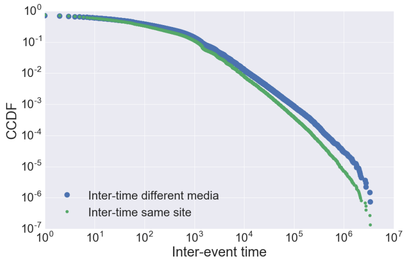


Fig. 2.18. **Inter-event time.** The distributions of the inter-event time between two successive posts published on different media (blue) and between two whatever successive posts (green).

social media. For instance, the sequence of events in Figure 2.17 results from the union of the activity of the user in six different social media. The example emphasizes two characteristics of the event sequences: *i*) singularly each event sequence keeps an high level of burstiness, i.e. the bursty behavior of the aggregated sequence is the union of bursty event sequences and; *ii*) a period of high activity in the aggregated sequence does not imply that the user is highly productive in each social media along the same period. In the example (see Figure 2.17) the user is initially active on BipTV only, then s/he begins to adopt the other social media. But the activity on Twitter ceases almost immediately and the other activities interchange like in Tumblr or Pinterest. The characteristic *i*) can be captured by the distribution of the inter-event time on each social media reported in Figure 2.19. Here, we extract the distributions from the social media with the highest number of posts. From the analysis of the distributions we infer two main insights: *a*) most of the distributions seem to follow a heavy-tailed law and *b*) each social media is characterized by a distribution that is far from the others. For example the probability of having a short inter-event time is higher in Pinterest, Delicious or LastFM w.r.t the other social media. To verify the observation *ii*) we analyze the distribution of the inter-event time between two consecutive events that happen on different social media, a quantity strictly related to the propensity of switching among the media. The distribution shown in Figure 2.19 is similar to the overall inter-event time distribution. The figure suggests that users alternate and mix their activities. The heavy-tailed distribution also indicates that the time to switch between different social media is likely to be short even if longer

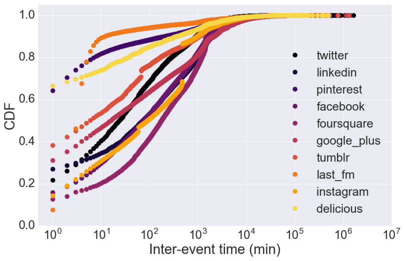


Fig. 2.19. **Inter-event time in the same social media.** The distributions of the inter-event time between consecutive posts on the same social media. The figure reports the distributions for the most used social media. For the sake of readability we use a linear scale on the y-axis.

periods are possible. This finding suggests that during a single day a user may publish on multiple social media. To this aim we introduce an index, the *post multiplexity* Ω, which captures the number of days a user is active on multiple social networks. Formally, let $p(u, s^i) =< p_0, p_1, \ldots, p_T >$ denote the sequence of posts published by the user u on the social media s^i, where p_i indicates the number of posts during the day i. We define a new binary sequence $p'(u, s^i)$, where $p'_i = \delta(p_i)$ [8] and the function $dm(u) = \sum_{s^i} p'(u, s^i)$ which computes the number of social media used in each day. The post multiplexity corresponds to the ratio between the number of days s.t. $dm(u) = 1$ and the overall number of sampling days. The index measures the propensity of a user of being multidimensional. A value close to 1 indicates that an individual uses only a social network per day. The distribution of the post multiplexity Ω has been reported in Figure 2.20. We consider different thresholds on the number of sampling days to make the results independent of the length of the sequence and reduce the effects of less active users. The distribution suggests that completely multidimensional users are quite rare as well as unmultiplex users; however most of the users tend to prefer a single media per day while sometimes they adopt multiple social media for posting.

[8] δ is the Heaviside function

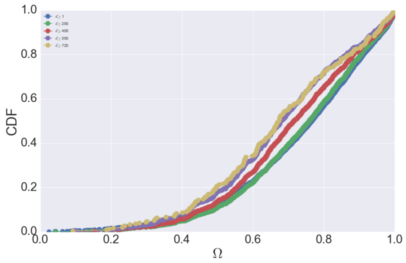


Fig. 2.20. **Posting multiplexity.** The distribution of Ω for different thresholds on the number of posts per user.

2.7 Patterns in username usage.

Besides providing information about people's behavior in using and posting across social sites, the collected datasets sum up properties concerning the choice of the username in different sites. The username represents the first information provided to the social site, the first way of being identified in the site and an element in the self-presentation. In fact, many behavioral aspects flow into the choice of the username, not least the limited memory capacity. On $D2$ we evaluate how users are coherent in the choice of the username by computing the edit distance and the complement of the Jaccard Index on the pairs (u_i^j, u_i^k) built from the set U_i of the usernames associated to the individual i. The former quantifies how dissimilar two strings are by computing the minimum number of operations to change u_i^j into u_i^k. The allowed operation are the insertion, the deletion and the substitution of a character with another. The latter computes the Jaccard coefficient on the sets of characters of u_i^j into u_i^k. In Figures 2.21 and 2.22 we report the distribution of the above measures computed on each possible pair and in the inner figures the same quantities computed on randomly shuffled pairs. In both cases we obtain a peak at 0, indicating exact matches or the usage of the same character set, while the randomization shift the peak of the distributions towards values typical of dissimilar strings. However a portion of the pairs indicates that people may change the username across social sites. An example is shown in Figure 2.23 where the change of the username between Pinterest and Google+ is more evident w.r.t. the average behavior reported in the previous figures. In fact, we observe that the decision of changing the username depends on the pair of social sites.

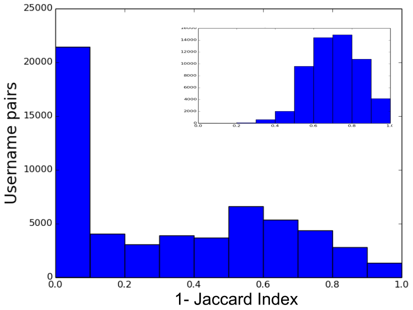


Fig. 2.21. **Distance distributions between usernames.** The distribution of the complement of the Jaccard Index between pairs of usernames associated to the same person.

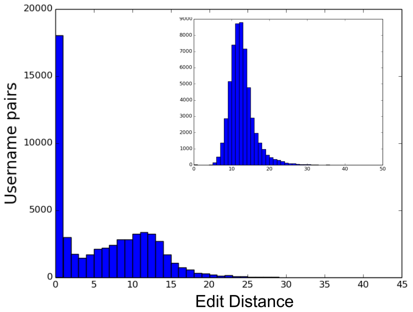


Fig. 2.22. **Distance distributions between usernames.** The distribution of the edit distance between pairs of usernames used by the same person.

2.8 Conclusion

To face the new challenge of giving an all-around picture of people's online behavior, in this chapter we perform an analysis on the same users across multiple social media. Our study relies on a new rich dataset gathered from the social media aggregator Alternion. It collects information about the way users post their favorite contents, their centrality on different social media and the usernames they choose. The study of this dataset led to the emergence of two main insights. On the one hand, we confirmed that the on-going phenomenon of the adoption of multiple social site is spreading. Not just people sign up in many social media but they are active and exploit their services as

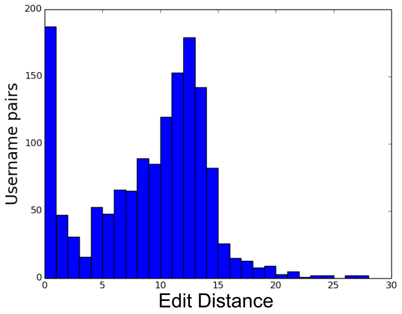


Fig. 2.23. **Google+ and Pinterest.** The distribution of the complement of the Jaccard coefficient computed on username pairs from Google+ and Pinterest.

well. On the other, the temporal information about how and when users post allowed us to investigate how people manage the opportunity of having different communication channels at their disposal. As far as we know, this study represents the first attempt to deal with the people's activities on multiple social media by using a large set of social platforms and users.

As regards the multiple adoption of social sites, the analysis of social media usage shows that the Alternion data captures the typical trend in today's users, despite the limitations discussed on the dataset section. The usage of multiple platforms is gaining momentum. In fact, half of the users signed up in at least three different social sites. This result stresses the importance of a multiplex approach when conducting studies which rely on online social networks. In fact, people are expressing their identity and their behaviors through multiple communication media. This observation is further strengthened by the results about active users. The fact that 73% of active users publish on at least two social sites means that people choose the right online channel to communicate and convey their contents. Also in this case the multiplex approach is fundamental in the extraction of people's interests and preferences. To this end, we plan to exploit these first results to study how users build their social identities across their social platforms [35]. In particular, we will verify whether different social norms characterize the major social platforms and if users adapt their self-presentation to these norms, as the results about the choice of the username may suggest. Finally, we will ask whether the identity of a single user expressed through the contents s/he publishes is persistent and coherent with the profile information.

The multiplex nature of the Alternion dataset allowed us to investigate the maintenance of users' popularity and centrality across social sites. In fact, despite the plethora of communication media, individuals are not confident with each media in the same way. For instance, someone better expresses and communicates through video, others through texts or blogs. That may result in different levels of engagement with fans or friends and, consequently, in different degree centralities. To this end, we asked whether a person's popularity is uniform across the social sites. The analysis led to not straightforward results and indicated that user's popularity in a given social site barely corresponds or does not correspond at all to his/her popularity on another social platform. Nevertheless, a more evident positive correlation between the posting activities across social platforms exists. In the future, we plan to analyze the reasons behind the weak correlations we observe. In particular, we will verify whether the social norms of the online platforms and the services they provide impact on the centrality of the users.

The multidimensional and longitudinal nature of the dataset is fundamental in the understanding of the human dynamics in communication and online social networks since it offers the opportunity to study how people handle different media. The primary goal of our analysis was to characterize the time-series associated to the posting activity of the single users across multiple social platforms and unveil their statistical properties focusing on measures which describe their level of burstiness. The analysis of the burstiness moved in two directions: i) we classified the aggregated behavior of each single user; and ii) within each user we characterized the bursty behavior in each single social media. The results show that i) singularly each event sequence keeps an high level of burstiness, i.e. the bursty behavior of the aggregated interaction sequence is the union of bursty event sequences and; ii) a period of high activity in the aggregated sequence does not imply that the user is highly productive in each social media during the same period. These observations represent the basis for a model of temporal patterns in multilayered social network which combines the idea of attention allocation and novel models which reproduce the human bursty dynamics [55]. We also plan to investigate the interaction sequences through frequent pattern analysis in order to highlight whether users are characterized by specific usage subsequences, i.e.

they have a predefined scheduling in the usage of their preferred social media. The above analysis will be combined to the content analysis of the posts to see if personal topics change along the observation period or users change their posting behavior in terms of covered subjects or whether some social platforms specialize on specific subjects.

Chapter 3

User Identification across Online Social Networks

Today, people spend part of their social life on the web, creating a virtual environment where they can find and meet friends, share and create information, and be engaged in a variety of social activities. The ability to gather the public traces left during these online activities would lead to a deeper understanding of the user's identity and behavior. This would improve service provisioning, enable service customization and cross-domain recommendation, and give rise to other services in different domains.

Even though we are dealing with information that users have explicitly designated as 'public', chasing them throughout different social networks is still a challenging research task. In fact, not all online services force users to specify their real identity, nor do they adopt platform-specific user's data (profile fields, friendships, tags), all which makes the field-matching difficult. In this scenario the identification of the same user across multiple social platforms would represent a key step, for it would facilitate the data gathering process.

How might the identification of a user be successful in practice when considering different social platforms? The answer heavily depends on the available data on the target OSNs. Accessing this kind of public data is not a straightforward task and gives rise to privacy concerns. To overcome these privacy issues, our proposed approach for identifying users across Google+/Facebook and Google+/Twitter relies on public data made available by their APIs. Among all fields returned by the APIs, we consider common fields only. This leads us to adopt a dynamic set of identification properties which flexibly adapts according to the systems under study, as opposed to the adoption of a pre-defined set like commonly assumed in literature (Carmagnola and Cena [15]).

In this chapter we propose an approach aimed at finding a match among identities across online social networks based on minimum common profile fields available through the APIs. We select the most effective features based on the common fields and then use automated classifiers to match the input profiles. According to the literature, in carrying out the identification task, random construction is the only way to construct negative instances, i.e. pairs of profiles corresponding to two different identities. The question which arises is: are random negative instances representative of the population w.r.t. training the classifier? To answer this question, we construct negative instances in three different ways and evaluate our trained classifiers on two different test sets. The accuracy and robustness of the approach are evaluated on two novel datasets collected from Google+, Facebook and Twitter. They consist respectively of 8,000 Google+/Facebook users and 2,400 Google+/Twitter users [1]. Moreover, we analyze the applicability of the learnt classifiers in a real scenario built on the Google+/Facebook users' neighborhoods. By comparing the proposed method with the approaches in Vosecky et al. [66]and Carmagnola and Cena [15], we show that our method is able to identify users with a significantly higher degree of accuracy.

3.1 Related work

Several database research studies and information retrieval communities have focused efforts on matching entities across different data sources [36], [67], [51], [4], [60]. While addressing similar problems, they do not match accounts directly across social networks.

In this section we summarize the studies related to the identification of individuals across different social media sites. We divide these studies into three different categories according to the information they rely on: username based identification, profile based identification and network and content based identification.

[1] The datasets have been made publicly available at `http://nptlab.di.unimi.it/?page_id=360`

3.1.1 Username based identification

Solutions based on username start from the assumption that the username is the minimum common factor available on several OSNs. Consequently, the different methods rely exclusively on features extracted from the strings composing usernames. Zafarani and Liu [70] proposed a methodology based on users' behavioral patterns when selecting their usernames. They demonstrate that the environment, the personality and the human limitations result in choices of the username that are not random but have redundancy. Identification features are constructed on endogenous and exogenous factors and on patterns due to human limitations. By using logistic regression, they obtained 0.930 and 0.927 of accuracy in the identification of the same user, before and after a feature selection step. To the authors' knowledge, Zafarani and Liu [70]work is one of the most complete research studies on this topic since their proposed features cover most of aspects of username creation.

The same authors [69] introduced a simple but interesting approach for finding other possible usernames that a user may select when s/he is registering on a social media. They empirically provided evidence on the possibility of identifying corresponding identities across various communities using usernames and a search engine. The approach starts by searching for a given username on Google to find a set of keywords that might represent possible usernames in the target social networks. Then this set is extended by adding/removing common prefixes and suffixes to/from its members. Finally, in order to filter out usernames, the existence of each username in the set is checked through the URLs that reside in the target community domain by searching on Google. An accuracy of 0.66 is reported. Since the accuracy of the approach depends heavily on the set of candidate usernames, its construction presents several challenges of its own. Furthermore the authors rely only on Google search results to determine the correctness of the identification.

Perito et al. [54] as well have explored the possibility of linking users' profiles simply by looking at their usernames. To link profiles that correspond to the same identity they estimate the uniqueness of a username by exploiting both a language model theory and Markov-Chain techniques. For each username the learnt binary classifier checks all possible usernames in a list for similarities. This makes the approach hard to use on a large scale.

Since the username is available in all social sites, username based identification approaches do not cope with challenges associated to the public information of users (e.g. heterogeneousness, incompleteness and falseness). However the exclusive usage of features extracted from usernames results in poor performances when two or more people have the same name or when users differentiate their usernames due to matters of privacy, etc.

3.1.2 Profile based identification

Carmagnola and Cena [15] addressed the subject of user identification for cross-system personalization among user-adaptive systems. They performed some analyses for defining a set of identification properties, the importance factors of each property and the relative thresholds. The proposed algorithm, which compares user profile attributes based on the assigned importance factors, is applied to users belonging to three user-adaptive systems developed in their research group. In the small test they used, 64 cases were positive matching ("identified") while 16 were negative ("not identified"). The results of the algorithm were respectively, "identified" in 59 out of 64 cases and "not identified" in 14 out of 16 cases.

Vosecky et al. [66] proposed a similar threshold based approach for comparing profile fields. They defined a weightingvector to control the influence of each profile attribute on the overall similarity. To compare user profile attributes they used exact, partial and fuzzy string matching and achieved an accuracy of 0.83.

Motoyama et al. [50] discussed a method for searching for and matching individuals on Facebook and MySpace by using profiles attributes. Their proposed method considers attributes as bags of words and calculates the similarity between two accounts as the number of common words between profile attributes. It does not account for common entities that have slightly different names.

3.1.3 Network and content based identification

Iofciu et al. [29] proposed an approach for user identification based on usernames and tags that users assigned to images and bookmarks. They also suggested various strategies for comparing the profiles of two users and were able to achieve an accuracy of about 0.6. Since matching accuracy via tags mainly depends on the number of tags assigned by a user, the accuracy of their identification rose from about 0.6 to about 0.8 by aggregating users' profiles from different sources.

Peled et al. [53] introduced an algorithm based on machine learning techniques to match two user profiles from two different OSNs. The classifiers utilized three types of features: name-based features, general user info based features, and topological based features including the number of mutual friends and mutual friends of friends of two users. Since the computation of the number of mutual friends strictly depends on the identification task, they just count the number of friends with identical names in both circles of friends. Their evaluation of the contribution of each feature in the classification process shows that the name-based features are the most important.

Goga et al. [24] conducted an investigation of the reliability of matching user profiles across real-world OSNs. They proposed a framework to understand how profile attributes used in the matching schemes affect the overall reliability. They used only public attributes, such as names, usernames, location, photos, and friends. These public attributes are not necessarily available through the APIs of the online social networks.

Jain et al. [30] proposed two identity search algorithms based on content and network attributes. They improved the traditional identity search algorithm founded on the attributes of the user profile. An average precision of 0.83 for the identity resolution process using profile, network and content identity search methods and an image-based identity matching method is obtained.

Malhotra et al. [46] applied automated classifiers for identifying profiles belonging to the same user across social networks. They used profile attributes and connections of users on different social networks to generate the digital footprints of the users. Their study indicates that user identification yields a large number of false positives.

Some other studies also leverage user connections to match accounts on social networks [38], [37], [68], [7].

The approaches in this category rely on data not retrievable by the APIs, so making them difficult to apply. For instance, Google+ API does not expose any resource to get the in/out neighbors, while Facebook requires the user's permission to access her/his friends.

Our approach belongs to the profile based category since it exploits fields common to the social media, in addition to the features extracted from the usernames. In Section 3.4 we compare our method with two of the profile based approaches. Moreover, in all the approaches based on classification techniques, the only way to construct negative instances is random selection; we show, however, that random construction in real scenarios yields a high rate of false positives. To overcome these limitations, we propose different methods for constructing negative instances. We do so by acting on the level of similarity. Lastly, we investigate the effectiveness of each approach for identifying a user across Google+/Facebook or Google+/Twitter.

3.2 Dataset

On most social networks users complete the basic information associated to their profiles by inserting links to other external web resources. Most of the time, these links refer to the different accounts users adopt in other social network sites. This behavior make possible the collection of the accounts associated to an individual person.

Operationally, we began by gathering profile information from Google+; in particular we exploit the technique in [25] to retrieve Google+ accounts. From the Google+ sitemap file, we randomly extracted more than one million user profiles. We analyzed the links referred to on each profile page, although only 2% of those users had indicated useful links. Although we limited our attention to links pointing to Facebook and Twitter, the method can easily include other social platforms. Depending on the social site a link pointed to, we retrieved the profile information through the specific API: namely, Facebook Graph API for Facebook and Twitter API for Twitter. The entire gathering process is depicted in Figure 3.1, where, as a last step, we select profiles containing Latin1 characters only. Finally, the different profile information associated with the same Google+ user are stored in the 'Profile DB', so that each record contains a sequence of profiles P^s, where s denotes the name of the social network.

Unfortunately, profiles in 'Profile DB' cannot be directly analyzed due to two main weaknesses: a) the data returned by each API are different from one another, in terms of field name and semantics (structural and semantic heterogeneity); b) some fields are empty or unavailable since users may not fill some common fields, may change a setting or may make some info private (data incompleteness). Both issues were analyzed. Results are reported in

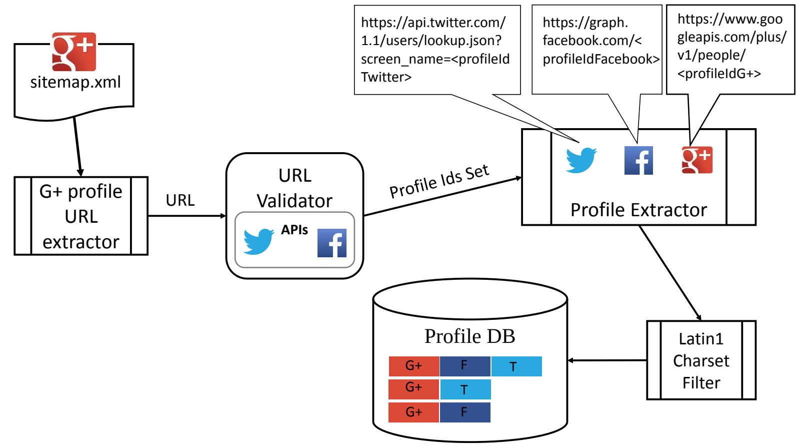

https://api.twitter.com/1.1/users/lookup.json?screen_name=<profileId Twitter>
https://graph.facebook.com/<profileIdFacebook>
https://www.googleapis.com/plus/v1/people/<profileIdG+>
sitemap.xml
G+ profile URL extractor
URL
URL Validator
APIs
Profile Ids Set
Profile Extractor
Profile DB
G+
F
T
G+
T
G+
F
Latin1 Charset Filter

Fig. 3.1. Data collection process. Once the links to other profiles have been obtained, the 'URL Validator' verifies whether the links are still valid and points to an account through the social service APIs. The 'Profile Extractor' gets the public information from each social site, while the 'Latin1 Charset Filter' detects the profiles containing Latin1 characters only. Each entry in the 'Profile DB' contains the profiles associated to each valid Google+ profile.

Figure 3.2. Regarding structural and the semantic heterogeneity, in Figure 3.2 we report — on the y-axis — all the fields returned by each social network API. Although the number and the meaning of the fields are heterogeneous across the social sites, for each pair of social sites we are able to identify some common fields that mainly involve data about username. The available common fields between Google+ and Facebook include username[2], first name, last name and gender. By contrast, in Google+ and Twitter profile common fields are username[3], full name, location and description.

To deal with data incompleteness, in Figure 3.2 we report the percentage of empty or missing fields grouped by social site. We observe that fields related to the name of the user — such as last, first and full name — contain information that is valid for most of the profiles in each social site. Secondly, even though

[2] In Google+ the username corresponds to the value associated to the field `displayName` of the `People` resource, while in Facebook the username corresponds to the field `username` in the `User` resource.

[3] In Twitter the username corresponds to the field `screen_name` related to the resource `users`.

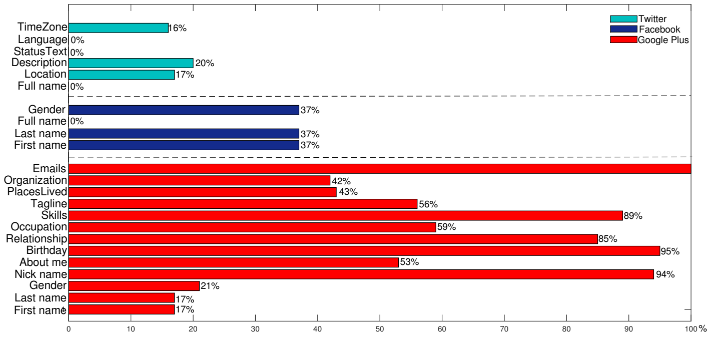


Fig. 3.2. Percentage of missing values in the fields of each social site.

the Google+ API offers more public fields than other social sites, its users tend neither to fill these fields nor keep them private. By contrast, on Twitter the same users are more prone to share information about where they live and what they are like.

We address the above issues by applying a data cleansing phase which removes the profiles with incomplete or missing data in the common fields. The process generates two datasets that we will use for building and testing the different user identification solutions:

- **Google+/Facebook dataset (GF).** GF dataset contains pairs of Google+ and Facebook profiles with common fields properly set. Initially the number of Google+ profiles linked to a valid Facebook account amounted to 14,000; however, after the cleansing phase, we discarded pairs (missing values included) and thus obtained 8,000 valid ones.
- **Google+/Twitter dataset (GT).** As with the GF dataset, this collection contains valid pairs of profile from Google+ and Twitter. Initially it consisted of 8,600 Google+/Twitter profile pairs, but after discarding pairs (missing values included) we ended up with 2,400 valid ones.

3.2.1 Username composition

We exploit the users' profiles to investigate how people in Google+, Facebook and Twitter use their real names to compose usernames. Our analyses are based on the first, last and full names provided by each user in her/is profiles. These analyses are helpful for selecting the most efficient identification features, in particular they help find redundant fields.

We identify four elements which contribute to the composition of the username: the first name, the last name and the full name, denoted as FirstN, LastN and FullN, respectively, while the fourth element (FullNnoSpace) is obtained by eliminating spaces from the full name. Then, we verify whether one of the above elements or their concatenations is/are substrings of the original username. Results are shown in Table 3.1.

One of the most remarkable results is that 100% of Google+ usernames contain the first name. We also observe that more than 45% of Facebook usernames contain no part of the real name (i.e. first name and last name). As expected, the most common way to generate Google+ usernames is by adding a space between the first name and the last name (97.91%). Since the Twitter API just returns the full name of users, the analyses on Twitter usernames do not include first name and last name. In our dataset 29% of Twitter usernames contain their full names. The results show that Google+ is considered as a more formal platform by users. After these empirical observations, in Section 3.3.2 we select the most appropriate features.

Substring	Google+(%)	Facebook(%)	Twitter(%)
$FirstN$	100	58.80	-
$LastN$	99.4	54.25	-
$FullN$	97.91	0.15	10.06
$FullNnoSpace$	0.52	9.97	29.19
$LastNFirstN$	0.96	1.05	-
$FirstN.LastN$	0	24.70	-
$LastN.FirstN$	0	0.76	-
$LastN < space > FirstN$	0.35	0	-
$\neg FirstN \wedge \neg LastN$	0	45.74	-

Table 3.1. Username composition. Basic substrings and their main concatenations contained in the username.

3.3 Methodology

In this section we present the methodology for solving the identification of the same identity across different social media. First we formalize the identification problem as a classification task. Then we illustrate how we build the feature set. We show that the construction of the negative instances in the generation of the training set plays an important role. In fact, we stress that the random construction of negative pairs results in a classifier which, in a real application incorrectly assigns a lot of profiles to a single user. To reduce the number of false positives we select negative instances by different methods. Finally, we propose different ways to construct training and test sets on top of which we evaluated different learning algorithms.

3.3.1 Problem definition

People leave plenty of various profile data and information on diverse OSNs. These fingerprints may be exploited to identify individuals across social networks and to integrate these sources to get an all-around vision of the users. However the amount of publicly available information about users' profiles is limited by the privacy setting and the APIs released by social media. Despite the limited availability of information, a common set of attributes exists for each pair of social sites. Usernames always represent the minimum common factor but, most of the time, gender, location and/or a short description of the person are made public by the APIs. Our methodology relies on these common attributes, which are available through social networks APIs, to identify users across OSNs.

In the profile identification we ask if given two profiles P^{s1} and P^{s2} from two different social sites $s1$ and $s2$, we can specify whether or not they belong to same individual. That corresponds to learn an identification function $f(P^{s1}, P^{s2})$ such that

$$f(P^{s1}, P^{s2}) = \begin{cases} 1 & \text{if } P^{s1} \text{ and } P^{s2} \text{ belong to the same identity} \\ 0 & \text{otherwise} \end{cases}$$

where the function $f(.,.)$ depends on the set of attributes common to P^{s1} and P^{s2}. An identification function can be learned using a supervised learning technique that employs selected features. As required by the supervised approach, the labelled data is based on the datasets into 'ProfileDB' (see Section 3.2), while the learning algorithms are the most adopted ones.

3.3.2 Feature extraction

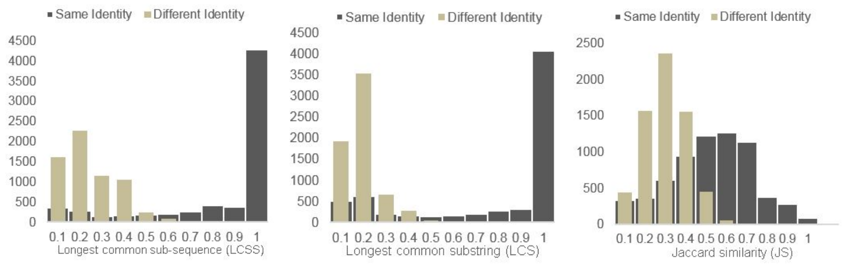
As in most of the learning frameworks, the choice of the appropriate features impacts the performance of the identification. Features are used as measures to indicate whether two profiles on different OSNs are similar in terms of behavioral patterns in the construction of usernames, in the composition of the short description and in the definition of the location. According to the available common profile fields through the APIs, most of the fields are string, so we adopt the main measures for string fuzzy matching. Specifically, we use the following metrics for comparing profile fields:

- **Exact Match(EM).** We use a string comparison function to check equality of data fields. We expect that this feature will be important for the identification; in fact, Zafarani and Liu [69] have shown that individuals tend to chose the same username.
- **Longest common substring (LCS).** Since adding prefix or suffix to a username, name and etc. is a common behavior across social networks. It can be detected by using LCS. We normalize this measure by dividing it by the average length of the two original strings to get a value in $[0, 1]$.
- **Longest common sub-sequence (LCSS).** We use the normalized longest common sub-sequence for detecting abbreviations.
- **Levenshtein distance (LD).** The Levenshtein algorithm (also called Edit-Distance) calculates the minimum number of edit operations that are necessary to modify one string to obtain another string [39]. It accounts for swapped letters, name shortening or the automatic composition of the username, e.g. the username suggested by Google+ or Facebook is a concatenation of the first name, a dot and the last name.
- **Jaccard similarity (JS)**: To compute alphabet overlaps, we use Jaccard similarity defined as the size of the intersection divided by the size of the union of the sample sets.
- **Cosine similarity with tf-idf weights**: The cosine similarity between two documents measures their similarity in terms of the angle between

their representation in the Vector Space Model, where for each term in the document set, we compute its tf-idf.

We employ exact match, longest common substring, longest common subsequence, Levenshtein distance and Jaccard similarity to compare name-based fields; otherwise, EM, LCS and JS are computed to evaluate the similarity of the "location" field. The "description" fields that users provide about themselves are first tokenized by removing punctuations and stop words; then we compute the cosine similarity between the two token sets. Finally, the exact match is applied to compare the "gender" field.

Figure 3.3 shows the distribution of some features for profiles pairs belonging to same identity. In the figure we also report the distribution of corresponding features for random pairs that do not belong to the same individual. These analyses are based on the GF dataset.



(a) (b) (c)

Fig. 3.3. (a) LCSS distribution of first name pairs. (b) LCS distribution of last name pairs. (c) JS distribution of username pairs

3.3.3 Training and test sets

Once the feature set is designed, the learning framework provides for the construction of training and test datasets to learn the identification function and then evaluate its performance in approximating the real one.

The datasets Google+/Facebook, GF, and Google+/Twitter, GT, only contain positive instances, i.e. pairs of profiles corresponding to the same identity. But we need negative instances to train the different binary classifiers.

In the literature, random construction is the means of choice for constructing negative instances. It includes creating negative instances by randomly producing pairs (P_i^{s1}, P_j^{s2}), such that P_i^{s1} is the profile of the user i on the social media $s1$ from one positive instance and P_j^{s2} is the j's profile on the site $s2$ from a different positive instance $(i \neq j)$ to ensure that they do not refer to the same identity. Consider two positive pairs (P_i^{s1}, P_i^{s2}) and (P_j^{s1}, P_j^{s2}), either (P_i^{s1}, P_j^{s2}) or (P_j^{s1}, P_i^{s2}) , to be constructed as negative instances in order to hold uniqueness of negative instances. If we consider both of these pairs as negative instances, the dataset would have a kind of redundancy with no positive effect regarding learning the function.

We applied Multilayer Perceptron and Random Forest learning algorithms to perform the classification task. The latter also assigns a probability to the instance to be evaluated, i.e. the probability that a given pair of profiles in different social sites refers to the same identity. We compared the learning techniques through the standard measures: accuracy, precision, recall and F-measure. Finally, the positive instances were randomly split into training/test sets with a 70:30 ratio, then we inserted the random negative instances. That results in training datasets TrainGF.1 and TrainGT.1 and in test datasets TestGF.1 and TestGT.1. The performances reported in Table 3.3 (TrainGF.1-TestGF.1 and TrainGT.1-TestGT.1 rows) show that both the Multilayer Perceptron and the Random Forest get very good performances for the identification task obtaining 0.95-0.96 as the F-measure.

Given the high precision and the low number of false positives, we tested our method in a real scenario. We applied the classifier trained on the training set obtained from the GF dataset to find the overlapping of the neighborhoods of users in Facebook and Google+. As representative of a general trend, here we focus on a random user who has 199 friends in her/his Facebook network and 112 persons in her/his Google+ network including her/his followers and followings. For each Google+ neighborhood of the random user we performed the identification task with respect to all her/his neighbors in Facebook. Namely, we want to discover common friends in Facebook and Google+. In our crawled dataset, for the Facebook neighborhoods of a random user only username and full name are available, while for her/his Google+ neighborhoods just usernames are available. We eliminated some features in our training set to fit the current instances fields. Since in this real application we do not have a ground-truth about the profile matching, we used the number of matched profiles for each candidate user as the main criterion for comparing and evaluating our classifier. As shown in Figure 3.4, most of Google+ friends have been matched with more Facebook friends. Specifically the average amount of false positives for the classifier trained on TrainGF.1 is 6. In general we found analogous results for most of the accounts we retrieved.

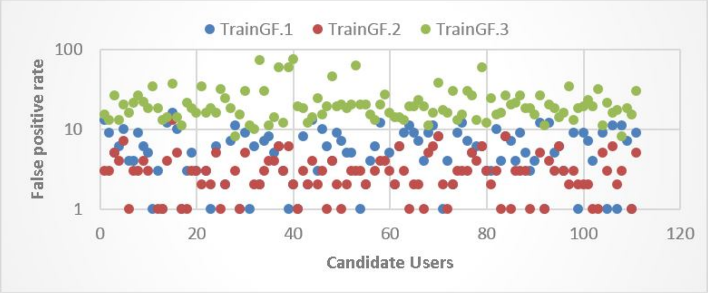


Fig. 3.4. False positive rate.

By analyzing TrainGF.1 we observed that, with the exception of fields like "gender" and "location", most of the random negative instances have completely different values for the corresponding fields. For example, in all 5,600 randomly constructed negative pairs based on the GF dataset, there were only 4 pairs with the same first name and one pair with same last name. However in a real application the identification algorithm must identify a candidate user in a group of users who usually belong to the same language, culture or region. In general the candidates show a high similarity/homophily resulting in a significant amount of similar fields. Therefore, applying a classifier trained on a dataset containing only random negative instances may result in a high rate of false positives. Thus, we propose two further methods to construct negative instances by taking into account these issues, as shown in Figure 3.5.

In the first method (Train2 in Fig.3.5), we build negative instances to provide a medium level of difficulty for the discrimination. To do this, 50% of negative instances are constructed randomly. The remaining 50% are set up in a way so that each negative instance has similar values at least in one field. This ensures that the number of negative instances with similar values are the same for each field. In the second method (Train3 in Fig.3.5) we construct negative instances to provide a more difficult setting. All negative instances

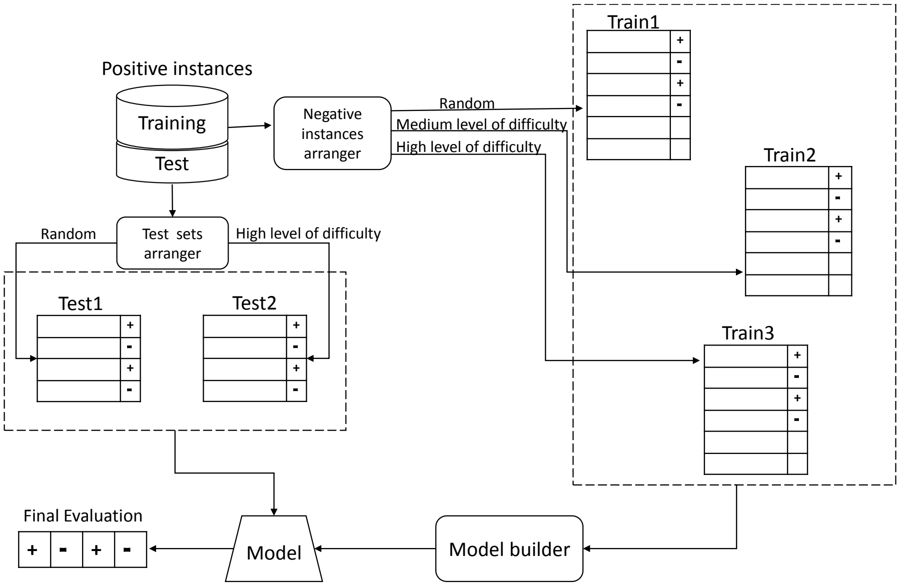


Fig. 3.5. The architecture of the proposed approach.

are set up in a way so that each negative instance has similar values at least in one field. This ensures that the number of negative instances with similar values are the same for each field.

We select our required instances from all possible negative instances that can be constructed randomly from positive instances. We also create two different test sets to evaluate the accuracy and robustness of our different classifiers. We separate 30% of positive instances for testing. Our first test set includes these positive instances and the same number of random negative instances. The second test set includes positive and negative instances that are difficult to discriminate; i.e., each positive instance has different values in at least one field and each negative instance has similar values in at least one field. Both test sets are balanced.

With respect to the identification between Google+ and Facebook profiles, we build the training datasets denoted as TrainGF.1, TrainGF.2 and TrainGF.3, respectively. All the training sets include 11,200 instances: 5,600 negative instances and 5,600 positive instances.

Whereas all the negative instances in TrainGF.1 are produced randomly, the negative instances in TrainGF.2 include 2,800 (50%) randomly constructed pairs, 1,400 (25%) pairs with same "first name" and 1,400 (25%) pairs with same "last name". 50% of pairs with same "first name" or "last name" have a same "gender": 1,400 pairs. We do not consider similarity in fields like "username", because we found no negative pairs with same "username" in all the possible randomly constructed negative instances. Moreover, personal information (name, gender, etc.) has a main role in selecting usernames by individuals. Therefore, negative instances with similarity in some personal information fields may have a kind of similarity regarding "username". Finally, negative instances in TrainGF.3 contain 2,800 (50%) pairs with the same "last name" and 2,800 (50%) pairs with the same "first name". Lastly, 50% of negative pairs in TrainGF.3 have a same "gender".

As for the construction of the test sets from the GF dataset, the first test set — TestGF.1 — includes 2,400 positive instances and 2,400 random negative instances. To create the second test set (TestGF.2), we select positive instances that are not equal in "firstname", "lastname" or "gender". Negative instances include (50%) pairs with the same "last name" and (50%) pairs with the same "first name", while 50% of negative pairs in TestGF.2 have the same "gender".

We apply the same procedure for the identification task between Google+ and Twitter. In this case 50% of negative instances in TrainGT.2 are built randomly. Finding instances with exactly the same value for the field like "Full name" is not possible since there is no match between the fields reporting first and last name. Therefore the remaining 50% negative instances are built in such a way so as to contain 33.3% pairs similar in the first part of the full name, 33.3% of instances similar in the last part of the full name and 33.3% of instances with same location. Also 50% of these instances have similarity in the "description" field bigger than the average cosine similarity of "description" field value in the all the positive instances. Finally, all negative instances in TrainGT.3 are constructed in a way so that each instance has similar values in at least one field.

To setup the test sets, the first testing set (TestGT.1) includes 50% positive instances and 50% random negative instances, while the second testing set (TestGT.2) includes positive instances that are different in at least one field. All negative instances are built so as to have comparable values in at least one field according to the same method adopted for constructing 50% of negatives instances in TrainGT.2.

3.4 Evaluation

Before evaluating the training sets on the different test sets, we validate our approach using 10-fold cross validation. As in the testbed application, we apply two different learning techniques: Multilayer Perceptron and Random Forest. Our aim is to verify if different learning algorithms can further improve the learning performance. These techniques have different learning biases, and so we expect to observe different performances for the same task. As seen in Table 3.2, results are not significantly different among these methods. When sufficient information is available in features, the user identification task is not sensitive to the choice of the learning algorithm. Using 10-fold cross validation, we get reasonably accurate results for both classification techniques on different datasets. We evaluate the effectiveness of each method in training the classifier on two test sets with different levels of discrimination. Table 3.3 shows the detailed results of applying the different learning techniques on the datasets.

Table 3.2. Results of the evaluation on our datasets using 10-fold cross validation.

Train	Random Forest				Multilayer Perceptron			
dataset	Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
TrainGF.1	97.38	0.974	0.974	0.974	97.19	0.973	0.972	0.972
TrainGF.2	93.48	0.937	0.935	0.935	93.21	0.935	0.932	0.932
TrainGF.3	91.63	0.921	0.916	0.917	91.59	0.922	0.916	0.917
TrainGT.1	94.82	0.950	0.948	0.948	94.22	0.944	0.942	0.942
TrainGT.2	91.27	0.916	0.913	0.912	91.35	0.917	0.914	0.913
TrainGT.3	91.63	0.919	0.916	0.916	90.76	0.909	0.908	0.907

As shown in Table 3.3, we confirm that the construction of negative instances in a random way is not a robust method. This is due to the fact that performances vary greatly from one test set to the other. The training on TrainGF.1 and TrainGT.1 results in the F-measure of 0.962 and 0.957 on TestGF.1 and TestGT.1 and in a worse F-measure of 0.551 and 0.844 on TestGF.2 and TestGT.2, respectively. TestGF.1 and TestGT.1 contain random negative instances, while TestGF.2 and TestGT.2 include pairs that are difficult to discriminate.

Table 3.3. Results of different classification techniques on the datasets.

Train dataset	Test dataset	Random Forest				Multilayer Perceptron			
		Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
TrainGF.1	TestGF.1	0.962	0.964	0.962	0.962	0.96	0.961	0.960	0.960
	TestGF.2	0.606	0.614	0.607	0.571	0.584	0.581	0.584	0.551
TrainGF.2	TestGF.1	0.926	0.933	0.927	**0.927**	0.927	0.936	0.928	0.927
	TestGF.2	0.796	0.840	0.797	**0.794**	0.765	0.787	0.766	0.765
TrainGF.3	TestGF.1	0.890	0.890	0.890	0.870	0.916	0.926	0.916	0.916
	TestGF.2	0.787	0.813	0.788	0.787	0.777	0.831	0.778	0.774
TrainGT.1	TestGT.1	0.954	0.955	0.954	0.954	0.957	0.958	0.957	0.957
	TestGT.2	0.847	0.848	0.847	0.845	0.844	0.850	0.845	0.842
TrainGT.2	TestGT.1	0.933	0.935	0.933	**0.933**	0.919	0.921	0.920	0.919
	TestGT.2	0.909	0.913	0.910	**0.910**	0.907	0.914	0.908	0.908
TrainGT.3	TestGT.1	0.897	0.898	0.898	0.898	0.901	0.901	0.901	0.901
	TestGT.2	0.917	0.925	0.917	0.917	0.911	0.913	0.911	0.911

A different behavior characterizes the classifiers trained through the second method (Train2) for building the training set. We observe the same patterns in the evaluation results on both GF and GT datasets. In fact, classifiers trained on datasets built based on the second method, TrainGF.2 and TrainGT.2, get the best F-measure on the test sets and exhibit more robustness, while classifiers trained on TrainGF.1 and TrainGT.1 show a high variation on the different test sets.

We also consider tests on unbalanced datasets. Table 3.4 shows detailed results on (75/25) and (25/75) unbalanced test sets. Although results on unbalanced test sets are close to results on balanced train and test sets, our second method shows better results also in the unbalanced setting. In fact, the average results on unbalanced test sets including 25% positive instances and 75% negative instances (TestGF.12, TestGF.22, TestGT.12, TestGT.22) are slightly better than results on balanced tests sets. The results on (90/10) and (10/90) unbalanced test sets are reported in Table 3.5. Again, they reveal the effectiveness of our second method versus the others.

As shown in Table 3.3, results are not significantly different between the two learning algorithms. In our experiments, Random Forest shows slightly better results in most of the cases; consequently, its results on balanced

Table 3.4. Results of different classification techniques on unbalanced test sets ((75/25) and (25/75)).

Train	Test	Distribution	Random Forest				Multilayer Perceptron			
dataset	dataset		Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
TrainGF.1	TestGF.11	75-25%	94.26	0.952	0.943	0.944	94.99	0.956	0.950	0.951
	TestGF.12	25-75%	97.80	0.978	0.978	0.978	97.41	0.974	0.974	0.974
	TestGF.21	75-25%	72.69	0.721	0.727	0.724	74.67	0.741	0.747	0.744
	TestGF.22	25-75%	59.84	0.769	0.598	0.623	60.60	0.772	0.606	0.630
TrainGF.2	TestGF.11	75-25%	89.88	0.927	0.899	0.904	89.26	0.924	0.893	0.898
	TestGF.12	25-75%	95.73	0.957	0.957	0.957	96.57	0.967	0.966	0.965
	TestGF.21	75-25%	75.32	0.857	0.753	0.770	74.67	0.860	0.747	0.764
	TestGF.22	25-75%	90.15	0.900	0.902	0.897	89.77	0.897	0.898	0.893
TrainGF.3	TestGF.11	75-25%	87.01	0.892	0.870	0.875	87.97	0.917	0.880	0.886
	TestGF.12	25-75%	87.80	0.891	0.878	0.882	95.95	0.961	0.960	0.958
	TestGF.21	75-25%	72.36	0.854	0.724	0.742	71.71	0.867	0.717	0.735
	TestGF.22	25-75%	88.63	0.884	0.886	0.881	90.90	0.911	0.909	0.904
- TrainGT.1	TestGT.11	75-25%	94.82	0.956	0.948	0.950	95.34	0.959	0.953	0.955
	TestGT.12	25-75%	95.68	0.957	0.957	0.957	96.20	0.962	0.962	0.962
	TestGT.21	75-25%	83.43	0.827	0.834	0.828	84.43	0.837	0.844	0.837
	TestGT.22	25-75%	60.96	0.797	0.610	0.632	67.10	0.815	0.671	0.692
TrainGT.2	TestGT.11	75-25%	92.24	0.938	0.922	0.925	90.34	0.920	0.903	0.907
	TestGT.12	25-75%	95.86	0.958	0.959	0.958	93.44	0.934	0.934	0.934
	TestGT.21	75-25%	90.41	0.923	0.904	0.908	89.22	0.915	0.892	0.897
	TestGT.22	25-75%	94.29	0.943	0.943	0.943	92.54	0.925	0.925	0.925
TrainGT.3	TestGT.11	75-25%	91.03	0.922	0.910	0.913	88.96	0.901	0.890	0.893
	TestGT.12	25-75%	91.37	0.921	0.914	0.916	88.62	0.897	0.886	0.889
	TestGT.21	75-25%	90.41	0.922	0.904	0.908	90.02	0.924	0.900	0.905
	TestGT.22	25-75%	94.73	0.948	0.947	0.948	95.17	0.951	0.952	0.951

datasets will be the reference method in the following experiments. Specifically, we analyze 1) how the classifiers, trained with different training sets, perform in the test bed application, and 2) whether our proposed method for user identification outperforms other methods in the literature.

3.4.1 Finding candidate users

We repeated the test bed experiment: namely, overlapping the neighborhoods, adopting the classifiers trained on the different sets. The results reported in Figure 3.4 show that the classifier trained on TrainGF.2 exhibits the best results, with three matches per neighbor on average, while the classifier trained on TrainGF.3 obtains the worst result with 21 matches per neighbor on average.

Since the Random Forest algorithm gives us the probability of a candidate username belonging to an individual, we can rank the candidates so as to verify

Table 3.5. Results of different classification techniques on unbalanced test sets ((90/10) and (10/90)).

Train	Test	Distribution	Random Forest				Multilayer Perceptron			
dataset	dataset		Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
TrainGF.1	TestGF.11	90-10%	93.86	0.962	0.939	0.945	94.26	0.961	0.943	0.948
	TestGF.12	10-90%	98.78	0.988	0.988	0.988	97.97	0.981	0.980	0.980
	TestGF.21	90-10%	76.19	0.821	0.762	0.789	78.57	0.841	0.786	0.811
	TestGF.22	10-90%	54.09	0.865	0.541	0.628	55.45	0.875	0.555	0.639
TrainGF.2	TestGF.11	90-10%	87.39	0.944	0.874	0.893	87.05	0.942	0.871	0.891
	TestGF.12	10-90%	98.58	0.986	0.986	0.986	98.45	0.984	0.985	0.984
	TestGF.21	90-10%	71.82	0.929	0.718	0.776	70.23	0.922	0.702	0.764
	TestGF.22	10-90%	94.09	0.940	0.941	0.940	94.09	0.940	0.941	0.940
TrainGF.3	TestGF.11	90-10%	85.97	0.938	0.860	0.882	85.63	0.939	0.856	0.880
	TestGF.12	10-90%	96.69	0.968	0.967	0.967	98.58	0.986	0.986	0.986
	TestGF.21	90-10%	67.85	0.927	0.679	0.745	65.87	0.926	0.659	0.729
	TestGF.22	10-90%	93.18	0.930	0.932	0.931	95.00	0.947	0.950	0.947
TrainGT.1	TestGT.11	90-10%	93.99	0.963	0.940	0.946	94.82	0.966	0.948	0.953
	TestGT.12	10-90%	97.30	0.975	0.973	0.974	96.48	0.970	0.965	0.966
	TestGT.21	90-10%	87.05	0.876	0.871	0.873	88.24	0.884	0.882	0.883
	TestGT.22	10-90%	61.05	0.883	0.611	0.688	62.10	0.893	0.621	0.696
TrainGT.2	TestGT.11	90-10%	91.09	0.953	0.911	0.922	89.85	0.947	0.899	0.912
	TestGT.12	10-90%	97.51	0.976	0.975	0.975	95.85	0.963	0.959	0.960
	TestGT.21	90-10%	88.96	0.939	0.890	0.904	87.76	0.936	0.878	0.895
	TestGT.22	10-90%	94.21	0.944	0.942	0.943	92.63	0.930	0.926	0.928
TrainGT.3	TestGT.11	90-10%	90.68	0.945	0.907	0.918	89.64	0.942	0.896	0.910
	TestGT.12	10-90%	90.89	0.944	0.909	0.919	88.19	0.932	0.882	0.898
	TestGT.21	90-10%	89.20	0.940	0.892	0.906	88.48	0.943	0.885	0.901
	TestGT.22	10-90%	94.21	0.947	0.942	0.944	95.26	0.950	0.953	0.951

whether or not the first positions contain the right profile. Specifically, for each user u in the Google+ neighborhood, let M be the set of neighbors in Facebook identified by the classifier and, for each node $m \in M$, let P_m be the probability that m is the same individual as u. We selected as a correct matching the m with the highest probability. Both the Google+ and Facebook neighborhoods were checked manually for overlapping users. Methods for manual checking were: checking for equal names, checking for similar profile pictures and other available information. In 87.75% of the cases, the right profile happened to be in first position by using TrainGF.2. This percentage falls to 10% and 82% when using TrainGF.1 and TrainGF.3, respectively. All experiments in this section show that the classifier trained on the dataset constructed through our second method (Train2) exhibits better performances in a real scenario.

3.4.2 Comparison with existing algorithms

The average performance of the classifiers trained on datasets based on the second method (TrainGF.2 and TrainGT.2) is better than the others. Now we need to compare the performance of the classifiers trained on these datasets with some acceptable approaches in the literature. Thus, we consider Vosecky et al. [66] and Carmagnola and Cena [15] methods for comparison. For implementing these approaches, we have to calculate the required parameters and thresholds on TrainGF.2 and TrainGT.2 and then evaluate the methods on the test sets.

In Carmagnola and Cena [15] each identification property is characterized by three parameters to calculate the importance factor of each property in the identification process: a) level of uniqueness (UL), which represents how much a property may assume the same value across different users. This property is directly related to the capability of identifying the user; b) values per user (VpU), representing the possibility for a feature, to be provided with different values for a unique user to the systems s/he interacts with and c) misleading level (ML), expressing the probability, for a feature, to be provided with a false value. VpU and ML are inversely related to the ability to identify the user. We report the values assumed by these three factors in Table 3.6.

Table 3.6. Calculated parameters for Carmagnola and CenaâĂŹs approach.

		Username	First name	Last name	Gender
TrainGF.2	UL	1.0	0.868	0.827	0.636
	VpU	0.999	0.224	0.243	0.021
	ML	0.520	0.218	0.242	0.056
		Username	Full name	Location	Description
TrainGT.2	UL	1.0	1.0	0.559	1.0
	VpU	0.997	0.534	0.771	0.995
	ML	0.493	0.343	0.477	0.493

After calculating the importance factor of each property by using UL, VpU and ML, we use the following formula to combine the importance factors values of all the matching properties.

$$IF = p + (1-p)q + (1-p)(1-q)m + (1-p)(1-q)(1-m)n \tag{3.1}$$

where p, q, m and n represent the importance factor of each identification property whose values match. The *IF* must exceed the importance factor threshold for the user to be considered as identified user.

Using the parameters and threshold of 0.82 calculated on TrainGF.2, the F-measure of 0.816 and 0.490 is obtained on TestGF.1 and TestGF.2, respectively. Moreover, by calculating the parameters and the threshold of 0.45 on TrainGT.2, the best F-measure of 0.659 and 0.602 is obtained on TestGT.1 and TestGT.2, respectively. The main weakness of the Carmagnola and Cena's approach is that it uses only the exact match as similarity measure. That also results in the lower accuracy on TestGT.1 than on TestGF.1. Moreover, since in our GT dataset almost half of the available fields are not name-based, including "location" and "description", the exact match is not a good choice for comparing these fields.

Since TrainGT.2 and TestGT.2 contain negative instances with similarity in the first part or last part of full names, better results are observed on TestGT.2 than TestGF.2. It can be observed that our method outperforms the method of Carmagnola and Cena by 0.111, 0.304, 0.274 and 0.308 on TestGF.1, TestGF.2, TestGT.1 and TestGT.2, respectively.

In Vosecky et al. [66] a similarity vector V is defined as $V(P^{s1}, P^{s2}) =< v_1, v_2, ...v_n >$, such that $v_i = comp_i(f_{i,P^{s1}}, f_{i,P^{s2}})$ where $f_{i,P^{s1}}$ is the i^{th} field of profile P^{s1} , for eachv_i, $0 \preceq v_i \preceq 1$, $|V| = |P^{s1}| = |P^{s2}|$.

For the purpose of the vector comparison algorithm, three categories of field matching are distinguished: exact matching, partial matching and fuzzy matching. A weight vector is defined to control the influence of each profile attribute on the overall similarity. In line with the Vosecky's method we selected the weights and thresholds on our datasets, as reported in Table 3.7. By calculating thresholds of 1.15 and 0.45 on TrainGF.2 and TrainGT.2, we

Table 3.7. Vosecky's weight vectors for our datasets.

Calculating weights on:	Username	First name	Last name	Gender
TrainGF.2	1.2	0.5	0.7	0.2
	Username	Full name	Location	Description
TrainGT.2	0.8	0.9	0.4	0.2

observe that the experiments on TestGF.1, TestGF.2, TestGT.1 and TestGT.2 show the best F-measure of 0.852, 0.535, 0.816, and 0.781, respectively. Our proposed method outperforms Vosecky et al's on the TestGF.1, TestGF.2, TestGT.1 and TestGT.2 by 0.075, 0.259, 0.117 and 0.129, respectively. As shown in Table 3.8, Vosecky's method exhibits better results if compared with Carmagnola and Cena's approach since the former exploits different similarity scores through the VMN function.

These experiments reveal that not only our proposed approach is robust and achieves good results on the different test sets, but also that it outperforms both Vosecky et al.'s and Carmagnola and Cena's approaches.

Train sets	Test sets	Carmagnola and Cena.	Vosecky et al.	Our approach
TrainGF.2	TestGF.1	0.816	0.852	**0.927**
	TestGF.2	0.490	0.535	**0.794**
TrainGT.2	TestGT.1	0.659	0.816	**0.933**
	TestGT.2	0.602	0.781	**0.910**

Table 3.8. Performance comparison of the profile based approaches.

3.4.3 Features Importance Analysis

Until now, we proposed different features measuring similarity between two fields. In this section we analyze how the feature's importance changes among datasets by using Information Gain Ratio.

The most import features in TrainGF.1 are all based on first name. This result shows the flaw of the random selection approach as the number of negative instances with same first name is too small to properly perform the learning task. In TrainGF.2 we constructed enough negative instances with the same first and last names. In this case, features based on other attribute like username is ranked higher in the list, which is a more reasonable result. The most important features in TrainGF.3 are based on username. In TrainGT.1, TrainGT.2 and TrainGT.3 , the most important features are all based on full name.

All these results show the importance of name-based features (username, first name, last name, full name) in the identification process. The five top important features on the datasets are presented in Table 3.9.

Table 3.9. Most important features using the Random Forest classifier.

	TrainGF.1	TrainGF.2	TrainGF.3	TrainGT.1	TrainGT.2	TrainGT.3
1	First names LCS	First names LCS	Usernames LCSS	Full names LCS	Full names LCSS	Full names LCSS
2	First names LCSS	First names LCSS	Usernames JS	Full names LCSS	Full names LCS	Full names LCS
3	First names JS	Usernames LCSS	Usernames LD	Full names LD	Full names JS	Full names JS
4	First names LD	First names JS	First names LCS	Full names JS	Full names LD	Full names LD
5	Last names LCS	Usernames JS	First names LCS	Usernames LCS	Usernames LCSS	Usernames LCSS

3.5 Conclusion

In this chapter we have proposed an innovative methodology for connecting people across Google+/Facebook and Google+/Twitter. Our method adopts minimal common information available through the official APIs of Facebook, Twitter and Google+ to drive features that can be used by supervised learning to effectively connect users across different online social networks. Relying on information available through APIs, our approach reduces privacy concerns as well as difficulties in collecting user information. By focusing on all common properties that characterize the systems under study, rather than on a predefined set of properties, we show that a better identification process can be achieved. It is one which simply uses available information across different platforms.

We constructed negative instances in three different ways, going beyond the commonly adopted random selection to evaluate the robustness of our identification algorithm on different datasets. Results show that the approach can lead to a very effective identification method and methodology for building reliable datasets. Moreover, we analyzed the success of our method in a real scenario built on Google+/Facebook neighborhoods.

Experiments show the advantages of the proposed method in comparison to previous methods; they also indicate that constructed features contain adequate information for connecting corresponding users.

Future work includes defining a framework for user identification across multiple online social networks and analyzing the benefit of connecting users in different domains.

Chapter 4

Effect of offline sociality on online interactions

Social network growth and evolution has been the subject of a wide literature in the last decade, but still it is an open problem and a very challenging, yet critical, issue. The fundamental process of network growth is link creation. Link creation mechanisms have been extensively studied and modeled by using features intrinsic to the network itself, shedding a light on the importance of common neighbors, triadic closure and homophily [3, 27, 58]. We are all aware and often arguing[56] that also the real life of users plays an important role in their online friendship creation. Online social networks, especially Facebook, born to be the mirror of human sociality and, at least partially, this still holds. Meetings and events are favorite ways to get new friends in real life. But the mechanism and the extent of the impact of offline meeting events on the creation of online friendships has not yet been studied, mainly because of the lack of available dataset.

More and more meetings and events are advertised on online social networks where people seek for their interesting events or are invited to an event. They announce their attendance some days before the event and participate the event physically on the day of the event. We leverage this functionality of Facebook to collect a dataset of events advertised on Facebook along with the information of nodes attending the event and temporally annotated new links between users interested in the event.

We take a first step towards understanding the effect of offline events on the graph structure of the social network where they are advertised. More precisely, we perform a temporal analysis of the *event social network*, constituted by people declaring to attend the event on Facebook and the links between them, and evaluate how it evolves during the event time period which is assumed to last from one week before to one week after the event occurs. We measure the network evolution from a global perspective by considering its communicability and evaluating whether the event increases the total communicability of the event network. From a 'per node' perspective, we evaluate how much the event helps people expanding their ego-networks by considering user's degree variation during the event time period. Since one common mechanism of link creation in social networks is triadic closure, to make in-depth-study, we compute the clustering coefficient and the number of new triangles, too.

We discover that offline events highly impact the online social networks, actually. To the authors' knowledge, there is no research addressing the issue of evaluating the impact of events on social networks like Facebook, but Liu et al [43] who studied a very special class of online social networks, the event-based social networks(EBSNs), only.

4.1 Dataset

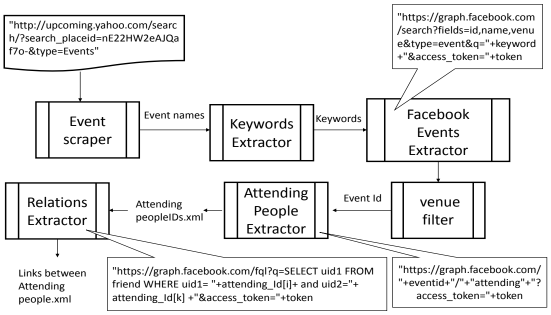

"http://upcoming.yahoo.com/search/?search_placeid=nE22HW2eAJQaf7o-&type=Events"
"https://graph.facebook.com/search?fields=id,name,venue&type=event&q="+keyword+"&access_token="+token
Event scraper
Event names
Keywords Extractor
Keywords
Facebook Events Extractor
Relations Extractor
Attending peopleIDs.xml
Attending People Extractor
Event Id
venue filter
Links between Attending people.xml
"https://graph.facebook.com/fql?q=SELECT uid1 FROM friend WHERE uid1= "+attending_Id[i]+ and uid2="+ attending_Id[k] +"&access_token="+token
"https://graph.facebook.com/"+eventid+"/"+"attending"+"?access_token="+token

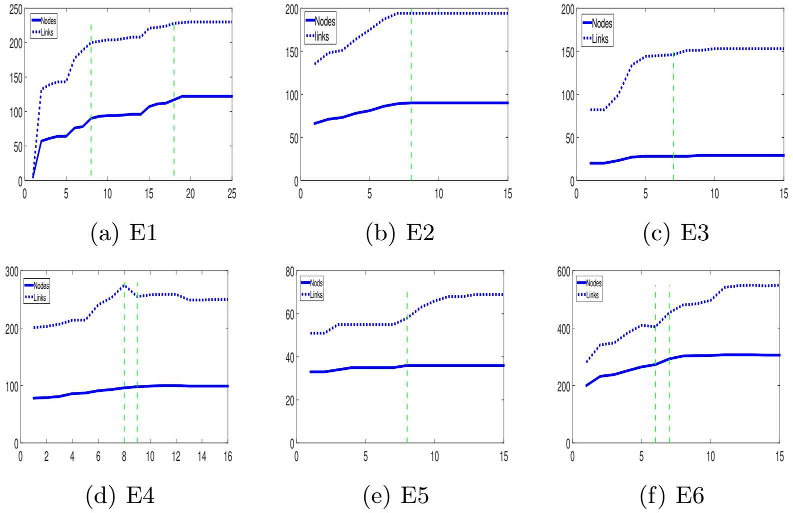
Fig. 4.1. Data collection procedure.

Using Facebook API and web scraping, we extracted a set of events that toke place in a European city [1]. To this aim, we developed a crawler to re-

[1] The visualization of the dynamic graphs of events can be found at the following link:https://www.dropbox.com/sh/00waikvphcu0dsh/AADVqFqSLUF-5YA7cIbZ9s5Za?dl=0

trieve the name of the events from Upcoming Yahoo web site[2]. We split each event name into its keywords in order to gather more events by querying each keyword from Facebook Graph API. The collected events whose venue field is not our target city are filtered out. After requesting the username of attending people from the API, it was possible to retrieve the friendship relations between them. Events with too few attending people are not considered in the following analysis. The detailed method of our data collection is illustrated in Figure 4.1.

The collected events range from cultural, art, musical and entertainment activities to informal get-together. The list of the events and their details are presented in Table 4.1. We consider the social network built by people attending the events and their friendship relations on Facebook and study its evolution from one week before the start date to one week after the end date. Note that we consider only people attending the events as nodes and only friendships between attendees as links. The number of nodes and links of the six events advertised on the social network at the end of the observation period is presented in Table Table 4.1, while their evolution over time is presented in Figure 4.2.



(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

Fig. 4.2. The evolution of the number of nodes and links over time for each event. Vertical dashed lines indicate the start and end date of the event.

[2] http://upcoming.yahoo.com

Table 4.1. Events advertised on Facebook and some properties about the related social networks.

Event ID	Event name	Start Date	End date	Number of nodes	Number of links
E1	Otello Party	5/9/13	5/19/13	122	230
E2	Image Art Fair	5/10/13	5/10/13	90	194
E3	Radio Live Concert	5/11/13	5/11/13	29	53
E4	Adam Carpet Live Concert	5/11/13	5/12/13	99	250
E5	Talent Scout Festival	5/18/13	5/18/13	36	69
E6	A Comics Festival	5/18/13	5/19/13	306	550

4.2 Results

We perform a temporal analysis of the event social networks given by all Facebook users attending the event and their links. Specifically, we investigate how the networks change at macroscopic level by means of the network communicability, at mesoscopic level by analyzing the clustering coefficient trend and at microscopic level by observing the increase of users' degree.

4.2.1 Communicability

To evaluate the impact of an event on the social network of its attendees from a macroscopic point of view, we analyze how its communicability changes over the event time period. In fact, in this very special kind of social networks, which are very limited and driven by a meeting, the most relevant feature is understanding how a piece of information might flow across the whole network and the communicability measures how easy is to send a message between two nodes.

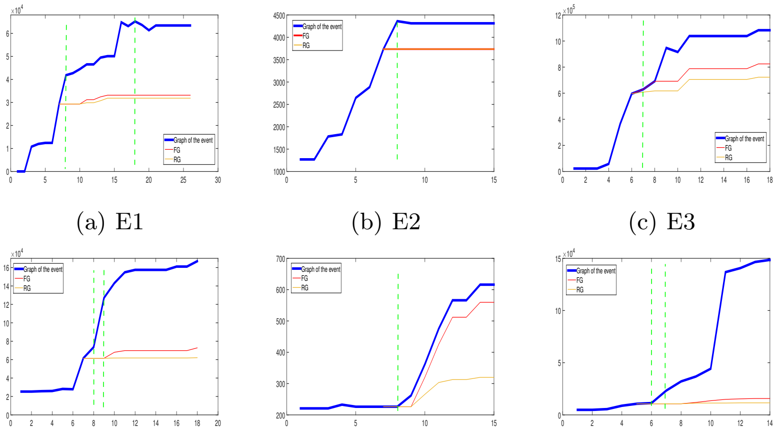
Many topological and dynamical properties of complex networks are defined by assuming that most of the contents on the network flows along the shortest paths. However, there are different scenarios in which non shortest paths are used to reach the network destination. Thus the consideration of the shortest paths only does not account for the global communicability of a complex network [19]. Estrada and Hatano [19] defined the communicability between two nodes by giving larger weights to the shorter walks and smaller weight to the longer walks. The communicability between two distinct nodes i and j is computed as:

$$C(i,j) = \sum_{k=1}^{\infty} \frac{(A^k)_{ij}}{k!} = [e^{\beta A}]_{ij}$$

where A is adjacency matrix of the network, and β is a tuning parameter. The normalized total communicability of a graph G including N vertices is defined as:

$$TC(G) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} C(i,j)$$

We compute the normalized total communicability (for $\beta = 1$) for the network of each event and for each time snapshot within a time window of one day. Figure 4.3 shows the evolution of the total communicability of all events. Vertical dashed lines indicate the start and the end dates of the offline event. The total communicability of all the events on our dataset increases over time with a similar pattern of growth. The communicability increases slowly initially, then increases exponentially, very similarly to the S-shaped growth curve. It rapidly increases from one day before the event, reaches its maximum in one to six days after it and then reaches a plateau. What is the reason of this growth pattern? The reason can be found or in new people joining the event or in old attendees creating new friendships among them. The former



(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

Fig. 4.3. The total communicability of the event graphs, FGs and RGs over time.

process has the effect of enlarging the event network, while the latter densifies it.

To distinguish between the two process, we observe how the communicability of a fixed group of people changes during the event. This way, we can isolate the effect of the network densification and check if it is the cause of the increase in the network communicability. Since the maximum increase in the number of nodes is mostly observed from the first day of the event period to the day before the offline event occurs, we consider as a fixed group (FG) the set of people present in the day before the offline event occurs when the maximum increase rate of attendees is usually registered.

Besides, we compare the communicability evolution between the network induced by FG and its random counterpart RG which is obtained by adding links randomly keeping fix the node set FG. Figure 4.3 shows the total communicability temporal behavior of the FG and RG networks in all events. For all the events, the communicability evolution of FG and RG networks are very similar and exhibit a small increase, if compared with the relevant growth observed in the global social event networks. Results show that an offline event advertised on Facebook enlarge attendees network and make easier the information spreading mainly because of new nodes and also for the creation of new links.

4.2.2 Clustering coefficient

At the mesoscopic scale, a network is described by clique and community. In the limited networks here considered, the most relevant constituents are triads which we measure by means of the clustering coefficient and the evolution of the number of triangles present in the network.

The clustering coefficient is the closeness of friend cliques in social networks [16]. The local clustering coefficient for an individual node i with K_i neighbors and $\mathit{\Gamma}_i$ edges between his neighbors is

$$c_i = \frac{\mathit{\Gamma}_i}{K_i(K_i - 1)}$$

while the total clustering coefficient is defined as:

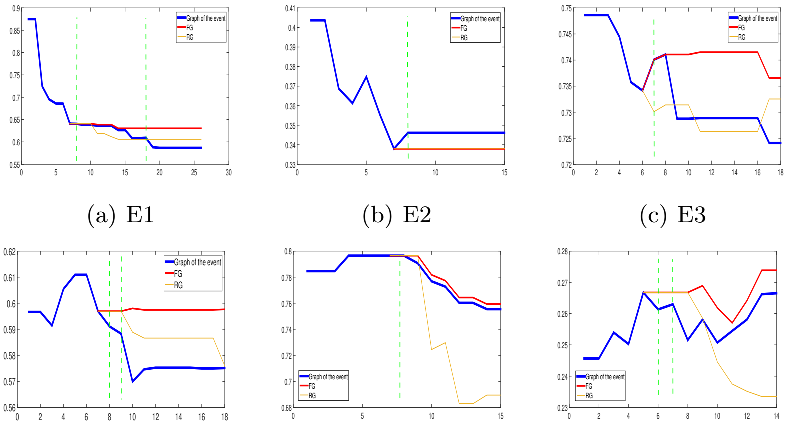
$$C = \frac{1}{N} \mathit{\Sigma} c_i$$

Kaiser [32] have shown that current definitions underestimate neighborhood clustering in a networks with many isolated or leaf nodes for which it is assumed: $c_i = 0$. Thus, instead of using N as the number of evaluated nodes for the global C, a new number N' indicating all nodes with defined local clustering should be used for a global measure C'. The relation between the new coefficient C' and the traditional measure C can be derived from the fraction of nodes that have one or zero neighbors, θ by

$$C' = \frac{1}{1-\theta}C \tag{4.1}$$

Since in the networks of events there is a considerable number of isolated and leaf nodes, we use 4.1 for calculating the clustering coefficient.

As shown in Figure 4.4, negligible variations over time are observed in the average clustering coefficient for all the events, but the first one which exhibits the maximum observed variation of 0.2. Figure 4.4 also shows the average clustering coefficient of FG and RG networks for all the events. The changes of the average clustering coefficient in the FG networks are also negligible, while in RG networks it decreases slightly more.



(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

Fig. 4.4. Average clustering coefficient of the events, FGs and RGs over time

To do more investigating on effect of event on number of triangles, one possible way is comparing number of triangles in first day of each event period to the last day. If the number of triangles in last day is higher than the first day it is more sensible to ascribe reason of this increase to new people joining the event whose friends already attend the event. Table 4.2 shows the number of triangles in first and last day of each event. However, changes in number of triangles in fixed group of nodes in the network of the event, with high probability is the effect of event. We count the number of triangles of FG and RG in the day before starting date and the last day. Table 4.3 shows the results. In all events number of triangles in the last day of FG is bigger than RG. All the results in this section reveals again this fact that new friendships are created during events and these new friendships close triangles.

Table 4.2. Number of triangles in the first day and the last day of each event.

	# of triangles(the first day)	# of triangles(the last day)
E1	2	216
E2	80	123
E3	163	438
E4	200	331
E5	26	51
E6	187	410

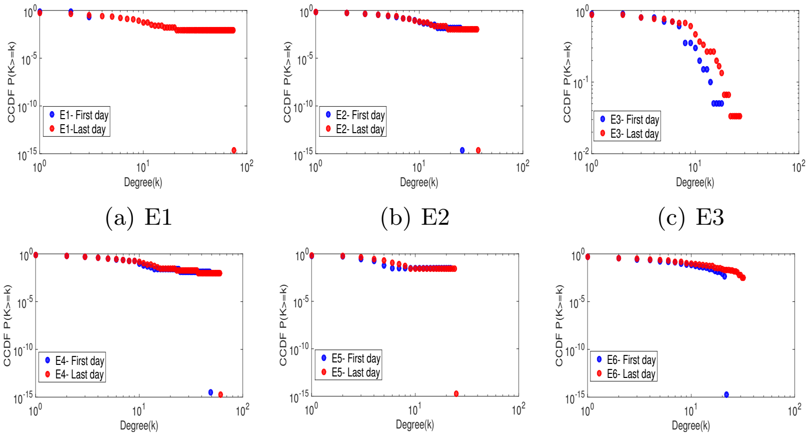
Table 4.3. Number of triangles of FGs and RGs in the day before starting date and the last day

	Number of triangles		
Event id	The day before Starting (FG,RG)	FG-The last day	RG-The last day
E1	158	164	162
E2	120	120	120
E3	382	410	386
E4	261	269	263
E5	34	51	39
E6	241	265	244

4.2.3 Degree

We compute the complementary cumulative distribution function (CCDF) of users' degree in the first and last days of each event period. As shown in Figure

4.5 all events in the last day include nodes with higher degree than the first day.



(a) E1 (b) E2 (c) E3

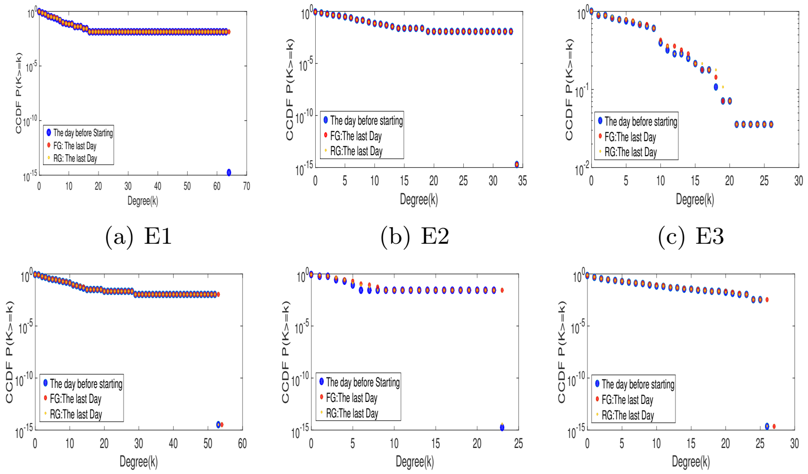
(d) E4 (e) E5 (f) E6

Fig. 4.5. CCDF of the degree.

But is this growth due to new users with higher degree attending the event or to old users who are expanding their ego-networks? To answer this question, we compute the CCDF of the users' degree in the FG and RG networks. The results are presented in Figure 4.6. For almost all events, the FG network in its last day includes nodes with degree higher than the one measured at the starting day of the observation and this effect is much more evident than in the random network RG.

We further consider the degree difference of each user between his first attendance and the last day of the event. The percentage of users expanding their ego-networks during E1, E2, E3, E4, E5 and E6 are 34%, 32%, 73%, 44%, 36% and 33%, respectively. The detailed results are summarized in Table 4.4. In this table, column one shows the event ID, while the second, third and fourth column present the percentage of users whose degree does not increase, increases between 1 to 10 and increases more than 10, respectively.

These results show that most of the nodes increase their degree during an event. To investigate whether this increase is due to new links between attendees or to new nodes who join the network, we consider the FG network as done for the communicability measure. We consider the degree variation of



(a) E1 (b) E2 (c) E3

(d) E4 (e) E5 (f) E6

Fig. 4.6. CCDF of the degree in FGs and RGs

Table 4.4. The percentage of users increasing their degree from their first day of online attendance to the last day.

Event ID	without increase	Between 1-10	More than 10
E1	66%	32%	2%
E2	68%	31%	1%
E3	27%	73%	0%
E4	56%	42%	2%
E5	64%	36%	0%
E6	67%	33%	0%

each nodes in FG between the day before the offline event starts and the last day. Detailed results are shown in Table 4.5 and Figure 4.7.

These results show that one of main reasons of the higher degrees observed in the last day of the events period is the creation of new friendships between the attending users. One remarkable result is that the degree of more than 80% of users with degree equal to 0 in FG remain zero during the event. That shows the important role of friends in creating new friendship in the network of events. One the other hand, as shown in figure 4.7, users with higher degree exhibit the greatest increase in the number of their friends.

Table 4.5. The percentage of users in FGs expanding their degree.

Event ID	Without increase	Between 1-10	More than 10
E1	67.11%	31.57%	1.32%
E2	94.19%	5.81%	0%
E3	50%	50%	0%
E4	72.53%	27.47%	0%
E5	68.57%	31.43%	0%
E6	72.70%	27.30%	0%

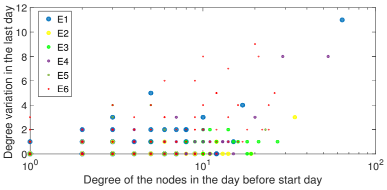


Fig. 4.7. Correlation between the degree of nodes in FGs the day before the start date and its variation.

4.3 Conclusion

In this chapter we take first steps towards understanding effect of offline events advertised on online social network on the graph structure of social networks. Some temporal analyses on events advertised on Facebook has been done. Communicability, clustering coefficient and degree variations for time windows of more than 15 days for each event was examined. We analyze the results and reveal the effect of events in new friendship creation between the attending people. The most activity and link creation take place during one day before event until one to six day after event. In future works by enlarging the dataset we could investigate the effects of events on the online social network focusing on specific domains such as culture and politics.

Chapter 5

Conclusion

The final Chapter concludes the dissertation by discussing and summarizing the main findings and the contributions. The main motivation underlying this dissertation conducted during this thesis is the increasing popularity of those social phenomena such as social media and, in particular, Online Social Networks (OSNs). In depth, in the present dissertation, we presented a comprehensive analysis on social networks, that are (i) the users' behavior across multiple social media sites.(ii) user identification across Online Social Networks (iii) the effect of users' offline meetings on their online friendship creation. In conclusion, we recall the main contributions of this dissertation:

(a) We studied in depth the human behavior across different online social networks on the new rich dataset that captures the typical trend in today's users.

The results of our analysis on the usage of multiple platforms stress the importance of a multiplex approach when conducting studies which rely on online social networks. In fact, people are expressing their identity and their behaviors through multiple communication media. This observation is further strengthened by the results about active users. The fact that 73% of active users publish on at least two social sites means that people choose the right online channel to communicate and convey their contents.

The dataset allowed us to investigate the maintenance of users' popularity and centrality across social sites. The analysis led to not straightforward results and indicated that a user's popularity in a given social site barely corresponds or does not correspond at all to his/her popularity on another social platform. Nevertheless, a more evident positive correlation between the posting activities across social platforms exists. Moreover, the dynamics of the online activity by a true multidimensional approach were evaluated. We found that the posting activity on online social media is bursty and highly heterogeneous.

This part of our work could be extended to analyze the reasons behind the weak correlations that we observed. In particular, it could be verified whether the social norms of the online platforms and the services they provide have an impact on the centrality of the users. Moreover, the interaction sequences through frequent pattern analysis could be investigated in order to highlight whether users are characterized by specific usage subsequences, i.e., they have a predefined scheduling in the usage of their preferred social media.

(b) we addressed the problem of user identification across online social networks by treating it as a classification task. We showed that using the standard approach to select negative instances in the literature results in a high number of false positives in practice. In fact, we stressed that the random construction of negative pairs results in a classifier which, in a real application, wrongly matches a lot of profiles to a single user. Three different ways for constructing negative instances were proposed to evaluate the robustness of our identification algorithm on different datasets. The results confirmed that the approach can lead to very effective identification method and methodology to build reliable datasets. Experiments revealed the advantages of the proposed method against previous methods and also indicated that constructed features contain adequate information for connecting corresponding users. Relying on minimum information available through APIs, the approach reduces the privacy concerns and the difficulties with collecting the users' information.

Our work could be extended by defining a framework for user identification across multiple online social networks and analyzing the benefit of connecting users in different domains.

(c) We took the first step towards understanding the effect of offline events on the graph structure of the social network where they are advertised. More precisely, we performed a temporal analysis of the *event social network*, constituted by people declaring to attend the event on Facebook and the links between them, and evaluated how it evolves during the event time period. The temporal analyses, including Communicability, clustering coefficient and

degree variations, revealed that offline events leave an impact on the online social networks and make easier the information spreading. The results showed that new friendships are created during events and that this new friendships creation is one of the main reasons of triangle closure and the higher degrees observed in the last day of the events period.

Our work could be extended by enlarging the dataset and focusing on the effects of offline events in specific domains such as culture and politics on the online social network. Moreover, the effect of the offline meeting could be investigated on the online relationship of different age groups.

References

1. Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, 23(2-3):169–209, 2013.
2. Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 835–844, New York, NY, USA, 2007. ACM.
3. Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):9, 2012.
4. Mishari Almishari and Gene Tsudik. Exploring linkability of user reviews. In *European Symposium on Research in Computer Security*, pages 307–324. Springer, 2012.
5. Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.
6. Matteo Barigozzi, Giorgio Fagiolo, and Giuseppe Mangioni. Identifying the community structure of the international-trade multi-network. *Physica A: statistical mechanics and its applications*, 390(11):2051–2066, 2011.
7. Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, and Hyungdong Lee. Joint link-attribute user identity resolution in online social networks. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM*, 2012.
8. Chris T. Bauch and Alison P. Galvani. Social factors in epidemiology. *Science*, 342(6154):47–49, 2013.
9. Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09. ACM, 2009.

10. Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Multidimensional networks: foundations of structural analysis. *World Wide Web*, pages 1–27, 2012.
11. S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. GÃşmez-GardeÃśes, M. Romance, I. SendiÃśa-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1 – 122, 2014.
12. Piotr Bródka, Przemysław Kazienko, Katarzyna Musiał, and Krzysztof Skibicki. Analysis of neighbourhoods in multi-layered dynamic social networks. *International Journal of Computational Intelligence Systems*, 5(3):582–596, 2012.
13. Charles D Brummitt, Raissa M DâĂŹSouza, and EA Leicht. Suppressing cascades of load in interdependent networks. *Proceedings of the National Academy of Sciences*, 109(12):E680–E689, 2012.
14. Francesco Buccafurri, Gianluca Lax, Serena Nicolazzo, and Antonino Nocera. Comparing twitter and facebook user behavior. *Computers in Human Behavior*, 52(C):87–95, November 2015.
15. Francesca Carmagnola and Federica Cena. User identification for cross-system personalisation. *Inf. Sci.*, 179(1-2):16–32, 2009.
16. Jing Cui, Yi-Qing Zhang, and Xiang Li. On the clustering coefficients of temporal networks and epidemic dynamics. In *ISCAS*, pages 2299–2302. IEEE, 2013.
17. Gregorio D'AGostino and Antonio Scala. *Networks of Networks: The Last Frontier of Complexity*. Springer Berlin Heidelberg, 2014.
18. Catherine Dwyer, Starr Hiltz, and Katia Passerini. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. In *13th Americas Conference on Information Systems*, AMCIS 2007.
19. Ernesto Estrada and Naomichi Hatano. Communicability in complex networks. *Phys. Rev. E*, 77:036111, 2008.
20. Sabrina Gaito, Gian Paolo Rossi, and Matteo Zignani. Facencounter: bridging the gap between offline and online social networks. In *Proceedings of the Complex Networks 2012 Workshop on Complex Networks and their Applications*, Complex Networks '12. IEEE, 2012.
21. Sabrina Gaito, Matteo Zignani, Gian Paolo Rossi, Alessandra Sala, Xiaohan Zhao, Haitao Zheng, and Ben Y. Zhao. On the bursty evolution of online social networks. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12, New York, 2012. ACM.
22. Qi Gao, Fabian Abel, Geert-Jan Houben, and Yong Yu. A comparative study of users' microblogging behavior on sina weibo and twitter. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, UMAP'12, pages 88–101, Berlin, Heidelberg, 2012. Springer-Verlag.

23. Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P. Gummadi. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, New York, 2015. ACM.
24. Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P. Gummadi. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1799–1808, New York, NY, USA, 2015. ACM.
25. Roberto Gonzalez, Ruben Cuevas, Reza Motamedi, Reza Rejaie, and Angel Cuevas. Google+ or google-?: Dissecting the evolution of the new osn in its first year. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 483–494. International World Wide Web Conferences Steering Committee, 2013.
26. László Gyarmati and Tuan Anh Trinh. Measuring user behavior in online social networks. *Network, IEEE*, 24(5):26–31, 2010.
27. Hong Huang, Jie Tang, Sen Wu, Lu Liu, and Xiaoming fu. Mining triadic closure patterns in social networks. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 499–504, 2014.
28. David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computer in Human Behavior*, 28(2):561–569, March 2012.
29. Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. Identifying users across social tagging systems. In *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '5. The AAAI Press, 2011.
30. Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. @i seek 'fb.me': Identifying users across multiple online social networks. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, 2013.
31. Hang-Hyun Jo, Raj Kumar Pan, Juan I Perotti, and Kimmo Kaski. Contextual analysis framework for bursty dynamics. *Physical Review E*, 87(6):062131, 2013.
32. Marcus Kaiser. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. 2008.
33. Marton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2, 2012.
34. Mikko Kivelä, Alexandre Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.
35. Farshad Kooti, Winter A. Mason, Krishna P. Gummadi, and Meeyoung Cha. Predicting emerging social conventions in online social networks. In *Proceed-*

ings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, pages 445–454, New York, NY, USA, 2012. ACM.

36. Hanna Köpcke and Erhard Rahm. Frameworks for entity matching: A comparison. *Data Knowl. Eng.*, 69(2):197–210, February 2010.
37. Nitish Korula and Silvio Lattanzi. An efficient reconciliation algorithm for social networks. *Proc. VLDB Endow.*, 7(5):377–388, January 2014.
38. Sebastian Labitzke, Irina Taranu, and Hannes Hartenstein. What your friends tell others about you: Low cost linkability of social network profiles. In *5th International ACM Workshop on Social Network Mining and Analysis, San Diego, CA, USA*, pages 1065–1070, 2011.
39. Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.
40. Huajiao Li, Haizhong An, Yue Wang, Jiachen Huang, and Xiangyun Gao. Evolutionary features of academic articles co-keyword network and keywords cooccurrence network: Based on two-mode affiliation network. *Physica A: Statistical Mechanics and its Applications*, 450:657 – 669, 2016.
41. Huajiao Li, Wei Fang, Haizhong An, Xiangyun Gao, and Lili Yan. Holding-based network of nations based on listed energy companies: An empirical study on two-mode affiliation network of two sets of actors. *Physica A: Statistical Mechanics and its Applications*, 449:224 – 232, 2016.
42. Huajiao Li, Wei Fang, Haizhong An, and LiLi Yan. The shareholding similarity of the shareholders of the worldwide listed energy companies based on a two-mode primitive network and a one-mode derivative holding-based network. *Physica A: Statistical Mechanics and its Applications*, 415:525 – 532, 2014.
43. Xingjie Liu, Qi He, Yuanyuan Tian, Wang-Chien Lee, John McPherson, and Jiawei Han. Event-based social networks: Linking the online and offline social worlds. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12. ACM, 2012.
44. M. Magnani and L. Rossi. The ml-model for multi-layer social networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11. IEEE/ACM, 2011.
45. Matteo Magnani and Luca Rossi. Formation of multiple networks. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 257–264. Springer-Verlag, 2013.
46. Anshu Malhotra, Luam Totti, Wagner Meira Jr., Ponnurangam Kumaraguru, and Virgilio Almeida. Studying user footprints in different online social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 1065–1070, Washington, DC, USA, 2012. IEEE Computer Society.

47. Pasquale de Meo, Emilio Ferrara, Fabian Abel, Lora Aroyo, and Geert-Jan Houben. Analyzing user behavior across social sharing environments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):14, 2013.

48. Giovanna Miritello, Esteban Moro, Rubén Lara, Rocío Martínez-López, John Belchamber, Sam GB Roberts, and Robin IM Dunbar. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, 2013.

49. Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

50. Marti Motoyama and George Varghese. I seek you: Searching and matching individuals in social networks. In *Proceedings of the Eleventh International Workshop on Web Information and Data Management*, WIDM '09, pages 67–75, New York, NY, USA, 2009. ACM.

51. Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, pages 173–187, Washington, DC, USA, 2009. IEEE Computer Society.

52. Vincenzo Nicosia and Vito Latora. Measuring and modeling correlations in multiplex networks. *Physical Review E*, 92(3):032805, 2015.

53. Olga Peled, Michael Fire, Lior Rokach, and Yuval Elovici. Entity matching in online social networks. In *Proceedings of the 2013 International Conference on Social Computing*, SOCIALCOM '13. IEEE Computer Society, 2013.

54. Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. How unique and traceable are usernames? In *Proceedings of the 11th International Conference on Privacy Enhancing Technologies*, PETS'11. Springer-Verlag, 2011.

55. Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Activity driven modeling of time varying networks. *Scientific reports*, 2, 2012.

56. Bernd Ploderer, Steve Howard, and Peter Thomas. Being online, living offline: The influence of social ties over the appropriation of social network sites. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 333–342. ACM, 2008.

57. Christian Quadri, Matteo Zignani, Lorenzo Capra, Sabrina Gaito, and Gian Paolo Rossi. Multidimensional human dynamics in mobile phone communications. *PLoS ONE*, 9(7):e103183, 07 2014.

58. Daniel Mauricio Romero and Jon M Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *ICWSM*, 2010.

59. Derek Ruths, Jürgen Pfeffer, et al. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.
60. Mudhakar Srivatsa and Mike Hicks. Deanonymizing mobility traces: Using social network as a side-channel. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 628–637, New York, NY, USA, 2012. ACM.
61. Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, and Jiawei Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11, pages 121–128. IEEE Computer Society, 2011.
62. Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.
63. Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. Rankclus: Integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 565–576, New York, NY, USA, 2009. ACM.
64. Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31), 2010.
65. Michael Szell and Stefan Thurner. Measuring social dynamics in a massive multiplayer online game. *Social networks*, 32(4):313–329, 2010.
66. J Vosecky, Dan Hong, and V.Y. Shen. User identification across multiple social networks. In *Proceedings of First International Conference on Networked Digital Technologies*, NDT '09. IEEE, 2009.
67. William E Winkler. The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer, 1999.
68. Gae-won You, Seung-won Hwang, Zaiqing Nie, and Ji-Rong Wen. Socialsearch: enhancing entity search with social network matching. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 515–519. ACM, 2011.
69. Reza Zafarani and Huan Liu. Connecting corresponding identities across communities. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, ICWSM '3. The AAAI Press, 2009.
70. Reza Zafarani and Huan Liu. Connecting users across social media sites: A behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13. ACM, 2013.

71. Reza Zafarani and Huan Liu. Users joining multiple sites. In *Proceedings of the Eights International Conference on Weblogs and Social Media*, ICWSM'14, Palo Alto, 2014. AAAI.
72. Reza Zafarani and Huan Liu. Users joining multiple sites: Friendship and popularity variations across sites. *Information Fusion*, 28:83–89, 2016.
73. Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.
74. Matteo Zignani, Christian Quadri, Sabrina Gaito, and Gian Paolo Rossi. Calling, texting, and moving: multidimensional interactions of mobile phone users. *Computational Social Networks*, 2(1):1–24, 2015.