University of Cyprus

DEPARTMENT OF COMPUTER SCIENCE

Online Social Networks Analysis: Extracting Insights and Evolution Trends

Hariton Efstathiades

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Cyprus

March, 2018

VALIDATION PAGE

Doctoral Candidate: Hariton Efstathiades

Doctoral Dissertation Title: Online Social Networks Analysis: Extracting Insights and Evolution Trends

The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at the Department of Computer Science and was approved on the ***March 28, 2018*** *by the members of the Examination Committee.*

Examination Committee:

Committee Chair _____
George Pallis

Research Supervisor _____
Marios D. Dikaiakos

Committee Member _____
Nicos Nicolaou

Committee Member _____
Elias Athanasopoulos

Committee Member _____
Ioannis Katakis

DECLARATION OF DOCTORAL CANDIDATE

The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

..... [Full Name of Doctoral Candidate]

..... [Signature of Doctoral Candidate]

Περίληψη

Οι πλατφόρμες κοινωνικής δικτύωσης (*Online Social Networks*) παρέχουν στον χρήστη τη δυνατότητα να αλληλεπιδρά, να εκφράζει και να μοιράζεται τις απόψεις, τα συναισθήματα και τα ενδιαφέροντά του με την οικογένεια, τους φίλους, τους συναδέλφους, τους γείτονες ή ακόμα και με άγνωστους. Αυτές οι αλληλεπιδράσεις δημιουργούν πολύτιμες πληροφορίες, καθώς οι χρήστες μπορούν πλέον να ενεργούν ως αναμεταδότες, μέσω των δημοσιευμένων μηνυμάτων τους, σε περιπτώσεις συμβάντων. Επιπλέον, με τη χρήση έξυπνων φορητών συσκευών και του Διαδικτύου, οι χρήστες είναι σε θέση να δημοσιεύουν πληροφορίες, ανεξάρτητα από τον τόπο στον οποίο βρίσκονται.

Η διατριβή αυτή παρουσιάζει ερευνητικές προσπάθειες στον τομέα της εξόρυξης γνώσης με τη χρήση δεδομένων που δημοσιεύονται στις πλατφόρμες κοινωνικής δικτύωσης. Αρχικά, ερευνούμε τη διαδικασία της συλλογής δεδομένων, καθώς η δημοτικότητα και οι τεράστιες ποσότητες πληροφοριών που δημοσιεύονται στις πλατφόρμες αυτές οδήγησαν στην καθιέρωση τους ανάμεσα στις κύριες πηγές δεδομένων σε διάφορους τομείς της ερευνητικής κοινότητας. Στα πλαίσια αυτής της έρευνας, σχεδιάσαμε, αξιολογήσαμε και παρουσιάζουμε ένα λογισμικό για την αποτελεσματική και αποδοτική συλλογή δεδομένων, το οποίο μας παρέχει μια τεράστια ροή πληροφοριών.

Η πρόσβαση σε αυτή τη δυναμική ροή δεδομένων των πλατφόρμων κοινωνικής δικτύωσης μας επιτρέπει να μελετήσουμε την εξαγωγή πληροφοριών σχετικά με το ανθρώπινο περιβάλλον, όπως είναι οι βασικές τοποθεσίες ενός χρήστη. Παρουσιάζουμε μια αποτελεσματική μεθοδολογία για τον εντοπισμό των βασικών τοποθεσιών ενός χρήστη, και συγκεκριμένα του τόπου κατοικίας και εργασίας. Η μεθοδολογία μας αξιολογείται με δεδομένα που συλλέχθηκαν από την Ολλανδία, το Λονδίνο και το Λος Άντζελες. Επιπλέον,

συνδυάζουμε τις πληροφορίες των πλατφόρμων *Twitter* και *LinkedIn* για την κατασκευή ενός συνόλου δεδομένων για την αξιολόγηση της μεθοδολογίας μας όσον αφορά την ανίχνευση της τοποθεσίας εργασίας. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη μεθοδολογία όχι μόνο υπερβαίνει τις τελευταίες τεχνολογικές μεθόδους κατά τουλάχιστον 30%, όσον αφορά την ακρίβεια, αλλά μειώνει και την ακτίνα ανίχνευσης τουλάχιστον στο μισό της απόστασης από άλλες μεθόδους. Για να αναδείξουμε τη δυνατότητα εφαρμογής της μεθοδολογίας μας και να παρακινήσουμε περαιτέρω έρευνα στην ανάλυση κοινωνικών δικτύων, αντιμετωπίζουμε τις πραγματικές προκλήσεις και συμπεραίνουμε ότι: 1) η δομή του γράφου, και 2) το συναίσθημα στις αλληλεπιδράσεις των χρηστών, σχετίζονται σε μεγάλο βαθμό με τις γεωγραφικές τοποθεσίες.

Τα αποτελέσματα σχετικά με την επίδραση των τοποθεσιών πάνω στο συναίσθημα μας οδήγησαν στο να μελετήσουμε περαιτέρω το πεδίο αυτό. Συγκεκριμένα, αναλύουμε το συναίσθημα μιας ειδικής κατηγορίας του εργατικού δυναμικού, τους νεοφυείς επιχειρηματίες (*Entrepreneurs*), και επισημαίνουμε τις διαφορές που έχουν με τον μέσο χρήστη του *Twitter*. Καταλήγουμε στο ότι υπάρχει συσχέτιση μεταξύ του συναισθήματος και της νεοφυής επιχειρηματικότητας.

Στη συνέχεια, στρέφουμε την προσοχή μας στην τοπολογία του γράφου του *Twitter*. Η πλατφόρμα από το 2015, αριθμεί πάνω από 500 εκατομμύρια χρήστες, από τους οποίους τα 316 εκατομμύρια είναι ενεργοί, δηλαδή συνδέονται στην υπηρεσία τουλάχιστον μία φορά το μήνα. Με τη μελέτη μας επανεξετάζουμε το δίκτυο που αναλύθηκε σε προηγούμενες ερευνητικές εργασίες, όπου αναλύθηκαν οι αλλαγές που παρουσιάζονται τόσο στο γράφο όσο και στη συμπεριφορά των χρηστών σε αυτό. Τα αποτελέσματά μας καταλήγουν σε ένα πυκνότερο δίκτυο, το οποίο δείχνει αύξηση του αριθμού των αμοιβαίων συνδέσεων, παρά το γεγονός ότι περίπου 12,5% των χρηστών του 2009 δεν ανήκουν πλέον στο *Twitter*. Ωστόσο, το μεγαλύτερο συνδεδεμένο στοιχείο του δικτύου φαίνεται

να μειώνεται σημαντικά, γεγονός που υποδηλώνει την κίνηση των συνδέσεων προς τους δημοφιλείς χρήστες. Επιπλέον, παρατηρούμε πολλές αλλαγές στις λίστες των χρηστών με μεγάλη επιρροή, έχοντας πολλούς λογαριασμούς που δεν ήταν δημοφιλείς στο παρελθόν να εξασφαλίζουν μια θέση στη λίστα των 20 κορυφαίων.

Abstract

Online Social Networking (OSN) platforms provide the user with the ability to interact, express and share her opinions, feelings and interests with the outside world; her family, friends, colleagues, neighbors or even strangers. These interactions hide much more valuable information, as users can now act as broadcasters in cases of events and incidents. Moreover, with the use of smart mobile devices and ubiquitous Internet connectivity a user is able to publish information, no-matter the place that she is located.

This thesis presents the research efforts towards providing knowledge using information published in OSN platforms. At first, we explore the area of OSN data collection as the popularity and huge amount of information published in OSN established them as one of the main data sources for a variety of research community fields. We design, evaluate and present a framework for efficient crowd crawling of Twitter, which provide us with an enormous stream of information.

The access to a highly dynamic OSN data stream, enable us to study the extraction of real-world information, such as the key locations of a user. We present an effective methodology for identifying a user's Key locations, namely her Home and Work places, and evaluate with Twitter datasets collected from the country of Netherlands, city of London and Los Angeles county. Furthermore, we combine Twitter and LinkedIn information to construct a Work location dataset and evaluate our methodology. Results show that our proposed methodology not only outperforms state-of-the-art methods by at least 30% in terms of accuracy, but also cuts the detection radius at least at half the distance from other methods. To illustrate the

applicability of our methodology and motivate further research in location based social network analysis we tackle real-world challenges and conclude that social graph structure and tweets sentiment are highly correlated with geographical locations .

The results on the influence of locations over sentiment trigger us to study further this field. In specific, we analyze the sentiment of a special category of workforce, entrepreneurs, and highlight the differences with the average non-entrepreneur OSN user. Our results suggest that there is a correlation between sentiment and entrepreneurship.

We then turn our attention on Twitter topology. The platform as of 2015, has more than 500 million users, out of which 316 million are active, i.e. logging into the service at least once a month. [1] With our study we revisit the network observed by Kwak et al. to examine the changes exhibited in both the graph and the behavior of the users in it. Our results conclude to a denser network, showing an increase in the number of reciprocal edges, despite the fact that around 12.5% of the 2009 users have now left Twitter. However, the network's largest strongly connected component seems to be significantly decreasing, suggesting a movement of edges towards popular users. Furthermore, we observe numerous changes in the lists of influential Twitter users, having several accounts that where not popular in the past securing a position in the top-20 list as new entries.

Hariton Efstathiades - University of Cyprus, 2018

[1] `https://about.twitter.com/company` (Last accessed: Jun. 2016)

Acknowledgments

I would like to express my gratitude to my thesis supervisors professor Marios D. Dikaiakos and George Pallis, for the guidance, trust and continuous support they have provided me during my Ph.D studies. I thank them for giving me the opportunity to work under their supervision, which was the trigger for my exposure to high-quality research. Special acknowledgments to my colleague and friend Demetris Antoniades for the insightful comments, motivation and knowledge he shared with me during my studies. I would also like to thank professor Nicos Nicolaou for his supervision on a part of this thesis.

Special thanks to Robert-Jan Sips and Zoltan Szlavik from IBM Benelux, for giving me the opportunity to work in a professional environment such as the IBM Center for Advanced Studies. I was very lucky for cooperating with them, gaining value from their background and experience. Moreover, I would like to acknowledge the support of Laboratory for Internet Computing members, with whom I had extensive fruitful discussions, received their feedback and exchanged ideas and insights during my studies. Additionally, I would like to thank all the colleagues who participate in iSocial, for their feedback and new insights during our research meetings and workshops.

Finally, I would like to thank my family for always being there for me when I needed them.

Thesis Contributions

This thesis is founded on the knowledge acquired by my involvement in the authorship of the following journal articles and conference papers:

Journal Articles

1. **"Sentiment of Entrepreneurs in Twitter"**, **H. Efstathiades**, J. Waters, N. Nicolaou, G. Pallis, and M. D. Dikaiakos, 2018. [*to be submitted*]

2. **"Users key locations in online social networks: identification and applications"**, **H. Efstathiades**, D. Antoniades, G. Pallis, and M. D. Dikaiakos, **Social Network Analysis and Mining (SNAM journal)**, vol. 6, no. 1, pp. 1–17, December 2016.

Conference Proceedings

3. **[Best Student Paper Award] "Online Social Network Evolution: Revisiting the Twitter Graph"**, **H. Efstathiades**, D. Antoniades, G. Pallis, M. D. Dikaiakos, Z. Szlavik, and R.-J. Sips, in **2016 IEEE International Conference on Big Data (IEEE BigData 2016)**, Washington, D.C, USA, December 2016

4. **"Distributed large-scale data collection in online social networks"**, **H. Efstathiades**, D. Antoniades, G. Pallis, and M. D. Dikaiakos, in **2016 IEEE International Conference on Collaboration and Internet Computing, (IEEE CIC 2016)**, Pittsburgh, PA, USA, November 2016

5. **"Identification of key locations based on online social network activity"**, **H. Efstathiades**, D. Antoniades, G. Pallis, and M. D. Dikaiakos, in Proceedings of the 2015 **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (IEEE/ACM ASONAM 2015)**, Paris, France, August 2015.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The development and growth of the World Wide Web has led to today's era, where the way information is generated, transfered and accessed has been radically improved. During last two decades an increasing variety of networking services has appeared and widespread across all the domains of our everyday lifes. This technological evolution and the increasing availability of access to different networking services has changed also the patterns of communication. Online Social Networks (OSN) entered our lives and became one of the main means of communication and interaction between people, as nearly everyone who actively accesses the Internet is also a user of at least one OSN platform, having 42% using more than one [11, 68]. Due to their user friendly interfaces and simple approach, different categories of users are attracted by these platforms. Their user directories are composed by users who belong to different demographic categories regarding their geographic locations, gender, age, education or profession.

OSN platforms provide the freedom to the user to build her profile with information that she choses to share, construct her community by choosing to connect with other users and interact with the latter by sharing different types of content. Due to this freedom of speech and flexibility, OSNs are an extremely valuable resource of information for several purposes. Data retrieved from these platforms are used to drive into conclusions or validate theories in a variety of fields, such as social science, market analysis and transportation [4, 34, 87].

Studying the content generated in Online Social Networking services became popular from the early years of their appearance. The nature and availability of this digital content enables new fields of study to arise and contribute significant findings to the research community. Several challenges have been introduced such

as understanding the different communities structure, identifying the influential factors in user's behavior and extracting knowledge from the different information.

This Ph.D thesis focuses on presenting the potentials and understanding on how OSN interactions could reveal useful real-world insights. We study through scientific methods the following problem domains: (*i*) Large-Scale Dataset collection from OSN, (*ii*) Extracting Key Locations from OSN activity, (*iii*) Influence of locations on OSN activity and mobility patterns, (*iv*) Sentiment of Entrepreneurs in OSN and (*v*) OSN evolution.

1.1 Thesis Motivation - Scope

At first we explore the area of OSN data collection as the popularity and huge amount of information published in OSN established them as one of the main data sources for a variety of research community fields. However, the design of a large-scale dataset collection campaign is a major problem for organizations and researchers who aim at addressing their research questions by analyzing this type of data. OSN platforms provide Application Programming Interfaces (API) to third party developers, which enable them to retrieve and use this data for application deployment. However, due to OSN imposed limitations, the process of retrieving large scale data with the use of these APIs is challenging and time consuming, resulting in datasets which are either incomplete or outdated. It is relatively impossible for an individual scientist or research group to follow an efficient dataset collection procedure and build a large sample in a short amount of time. With this work we present a framework for efficient crowd crawling of OSN. Our framework is based on the use of multiple OSN accounts, which are engaged in an efficient distributed collection process able to circumvent the imposed limitations without violating the terms of use. We present an evaluation of the proposed solution and demonstrate its performance in terms of dataset completeness and timeliness, for the case study of Twitter, one of the most popular platforms used in research.

The access to a highly dynamic OSN data stream, enable us to study the extraction of physical-world information, such as the key locations of a user. We present an effective methodology for identifying a user's Key locations, namely her Home and Work places, and evaluate with Twitter datasets collected from the country of Netherlands, city of London and Los Angeles county. Furthermore, we combine

Twitter and LinkedIn information to construct a Work location dataset and evaluate our methodology. To illustrate the applicability of our methodology and motivate further research in location based social network analysis we provide an initial evaluation of three such approaches, namely (*i*) Twitter user mobility patterns, (*ii*) Ego network formulation and (*iii*) Key location influence on tweets sentiment.

The results on Key locations influence over the sentiment trigger us to study further this field. In particular, we analyze the sentiment of entrepreneurs in OSN, and highlight the differences with the average non-entrepreneur user. The importance of emotion in entrepreneurship is recognized in a growing body of literature, and links have been found between an entrepreneur's emotions and various aspects of their cognition and behaviour that are central to the decision to establish companies. We start by arguing that entrepreneurship brings entrepreneurs freedom to set their own working conditions and objectives, which raises their general sentiment relative to non-entrepreneurs. However, we propose that entrepreneurs face specific threats associated to the establishment of new ventures, which lower their sentiment directed towards business matters relative to non-entrepreneurs. For social entrepreneurs, we argue that they may experience raised sentiments both due to altruistic enjoyment of other people's improved circumstances, as well as a "warm-glow" from personal participation in socially-oriented work. For serial entrepreneurs, we introduce arguments that entrepreneurial experience can lower people's sentiment due to memories of adverse events, as well as due to less pleasure being derived from novel experiences.

From our study we conclude that geographical locations are highly correlated to social graph structure. Thus, we turn our attention on Twitter topology. In 2010 the popular paper by Kwak et al. [58] presented the first comprehensive study of Twitter as it appeared in 2009, using most of the Twitter network at the time. Since then, Twitter's popularity and usage has exploded, experiencing a 10-fold increase. As of 2015, it has more than 500 million users, out of which 316 million are active, i.e. logging into the service at least once a month. [1] With our study we revisit the network observed by Kwak et al. to examine the changes exhibited in both the graph and the behavior of the users in it.

[1] `https://about.twitter.com/company` (Last accessed: Mar. 2018)

1.2 Approach and Methodology

The procedure of extracting knowledge from OSN platforms involves different supporting parts. For the purpose of this thesis we divide the workload in different pillars. Each of these pillars can be seen as a different problem domain of its own that justifies for a comprehensive investigation.

1.2.1 Large-Scale Dataset Collection From OSN

OSN platforms, such as Twitter, Facebook and LinkedIn, represent information in similar abstractions. Each user has a unique identifier, usually of type *long*. Similar identifiers are assigned in messages posted by the user, like *Tweets* and *Posts* in Twitter and Facebook respectively. The connections between the users are retrievable as edge lists, which denote either a reciprocal (friend) or direct (follower) connection between two users. The retrieval of this information can be achieved either using OSN provided API or through Web Scraping. However, the terms-of-services of popular OSN prohibit the data collection through automatic Web Scraping [2] [3]

We design and introduce an OSN data collection framework that addresses the current challenges and provides the end-user with large scale crowd crawling capabilities. Our proposed framework enables the performance of large scale dataset collection campaigns in the most efficient way, compared to state-of-the-art. We implement a crowd crawling prototype, using Twitter, and demonstrate its performance, with respect to OSN's request limiting policies.

A data collection campaign can be either *i*) Resource Specific collection or *ii*) a Real-Time stream collection. The first case provides retrieval services for a specific resource (e.g a user's profile, a tweet or post), while the latter enables the sample collection of real-time information that is being published in the OSN. Our framework provides two different services and enables both cases of a data collection campaign, which are 1. Resource Specific Data Collection, and 2. Real-Time Stream Collection. The proposed system uses a Map-Reduce-like approach to overcome several limitations and be able to run small instances for performance objectives.

[2] `https://twitter.com/tos`, Twitter Terms of Service (Last Accessed: March 2018)

[3] `https://www.facebook.com/apps/site_scraping_tos_terms.php`, Facebook Terms of Service (Last Accessed: March 2018)

1.2.2 Extracting locations from OSN activity

For the identification of a Twitter user key location we propose a methodology which uses geo-tagged data. We used a variety of OSNs to collect geo-location information about the users. It is based on two main observations regarding user's real life habits. These are: (i) users tend to spend a significant, but distinct, amount of their time during an average day in two key locations namely their Home and Work; (ii) these two locations are much more likely to appear in the user's geo-tagged activity during these specific timeframes, than locations that are not so frequent in user routine. We apply this methodology to identify the *Home* and *Work* location of the users. Evaluation of our method, using data from several geographical regions, showed that it outperforms previous methods by more than 30%. Additionally, it can identify the user's key locations at post-code granularity, that is in a radius smaller than 3Km. Comparison with socio-economic open data showed that our method can correctly identify the populated areas of the geographical region of interest.

To further evaluate our proposed methodology we illustrate how one can combine information from multiple social networks, namely LinkedIn and Twitter, in order to construct a dataset that includes both the user's work location and her tweet activity. Using this dataset we evaluated our method for work location identification. Our results show an accuracy close to 80% for identification of user location in a 10Km proximity. To the best of our knowledge this is the first attempt to construct a workplace ground truth dataset and also the first workplace identification method.

1.2.3 Identifying the influence of locations on OSN activity and mobility patterns

We then turn our attention on examining the influence of Key Locations in user's OSN activity and mobility patterns. We use the resulted dataset from the previous study and enrich it with open data from the 3 different geographical regions. This action enables us to identify the influence of home and work locations in users' daily mobility patterns. Additionally, we are now able to investigate how users' ego-networks are formed based on their key locations.

Furthermore, we aim on measuring the influence of locations on the mood that users express in their OSN publishing activity. Sentiment is commonly used to

measure the emotions of user's natural language. It can show the reaction of users to several events or their emotional state during a conversation. Combined with location information it can show how different geographical areas react to specific events or express themselves during their everyday online interactions. For example, Hedonometer is a tool used to measure the average happiness of Twitter users, also segmenting the tweets to the different US states they originate from [4]. Our method, able to identify the actual Home and Work locations of the users, can be used to zoom-in into the different neighborhoods and examine the sentiment of the different Key locations of the users.

1.2.4 Sentiment of Entrepreneurs in Twitter

Our results highlight the influence of Key Locations to the sentiment of the content that a user publishes. Based on these finding we study further this field, by identifying correlation between different types of users, such as *entrepreneurs*, and their sentiment. For this study we use datasets that we have already collected for the parts presented previously. Our database contains more than 13 billion messages of 140 characters or less, sent using the Twitter microblogging platform. We take the text written by a Twitter user as the principal source of our information for determined and determining variables, with other control variables concerning the tweet and the entrepreneur also extracted from Twitter.

We consider tweets sent by both entrepreneurs and non-entrepreneurs. Entrepreneurs are defined here to be Twitter users who have in their Twitter profile description any of the following terms: entrepreneur, founder, co-founder, business-owner, business owner, start-up, or start-up. We then construct the non-entrepreneurs datasets, with the same quantity of randomly sampled users. The data is then analyzed using sentiment analysis, a form of textual analysis, using NLTK. We calculate and evaluate our results using different statistical frameworks in R statistical language.

1.2.5 OSN evolution - Revisiting the Social Graph

Our analysis is based on two different snapshots of the same Twitter network: (*i*) the complete Twitter 2009 graph, as collected and shared by [58], and (*ii*) the collection

[4]Hedonometer, `http://hedonometer.org/index.html` (Last accessed: June 2015)

of the same list of Twitter users and their social graph as it appeared in late 2015. The 2009 graph was made available by Kwak et al.[5] According to the authors, the dataset represents the complete social graph of Twitter in 2009. Using the list of Twitter users that appeared in *TW2009* we perform a large-scale collection, through the current version of the Twitter API[6], with compliance to platform's terms of use and users' privacy.

Through this collection we retrieve the same set of Twitter users and their ego-network state (followers and followings) in November 2015. From this network we remove any connections (edges) that are directed towards or coming from users who do not belong in 2009 set. Thus, our *TW2015* snapshot contains only the connections that existed and have arise between the users that consisted the Twitter social network in 2009. In addition to the two full graphs of the 2009 Twitter users we also examine and compare, where relevant, the 2009 graph as it would appear if the users that belong to the above three categories where not existent in 2009.

1.3 Thesis Statement and Contributions

In this thesis we argue and demonstrate how, *sourcing OSN data properly and timely, can potentially lead to explanations of human behavior in relation to physical (offline) world, such as mobility patterns, interaction habits and emotion state.*

The contributions of this thesis can be summarized as follows:

- We design and present a crowd crawling data collection framework, which enables an individual or a group of researchers to efficiently perform large scale data collection campaigns with the participation of OSN users. The proposed solution is able to efficiently: *a*) Collect historical data in an asynchronous manner, *b*) Retrieve the OSN stream in real-time. We implement a proof-of-concept prototype, which demonstrates the system under a case study on Twitter. We present an extended evaluation on different types of devices along with a comparison over the state-of-the-art OSN data collection methods. We evaluate the proposed framework in both the large scale asynchronous data collection procedure and the collection of the real-time Twitter activity. Experimental

[5] `http://an.kaist.ac.kr/traces/WWW2010.html` (Last accessed: Jun. 2016)

[6] `https://dev.twitter.com/rest/public` (Last accessed: Jun. 2016)

results show that the proposed solution provides improvements in both data collection functionalities by more than 100x and 3x respectively.

- We present an effective method for identifying Key locations of a user based on geo-tagged Twitter data. The extended evaluation of our method shows that it can identify the user's Home location with an accuracy of more than 80%, giving a 30% improvement over the state-of-the-art. We construct a Work location identification dataset by using user reported information to both Twitter and LinkedIn OSNs and present an evaluation of user workplace identification with an accuracy of 63% at post-code level and more than 80% for a radius of 10Km. Furthermore, our results show that our method outperforms the state-of-the-art in terms of accuracy, identifying 90% of user's Work locations in a radius smaller than 20Km, compared to the 50Km radius needed from other approaches to reach similar levels of accuracy. To the best of our knowledge this is the first study that constructs a dataset and performs analysis for workplace location identification. We then use the proposed method to perform a broader Key location identification for all users in our dataset and compare that with socio-economic open data for the areas of interest. The comparison shows a clear mapping between our identified locations and the ground truth.

- We examine a number of applications of our method showing that users' Key locations can be used to identify Twitter user behavior both in terms of mobility and sentiment. Our findings show that users tend to live and spend their free time in close proximity to their Work location during weekdays. During weekends, users leisure travel distance increases to locations further away from their Home location. Moreover, our sentiment analysis results show that users tend to be far more positive in their tweeting behavior when tweeting from leisure locations rather than tweeting from their Home or Work locations. We show that a user's Ego network is mostly formed by users in close proximity to her Home location. Furthermore, the user's stronger connections, as defined by reciprocity, are not only in close proximity to the user but also in areas with similar economic status. Our findings show the impact of the real-world underlying social network in the formation of a user's Online Social Network.

- We examine the sentiment of traditional, social, and serial entrepreneurs, and

compared them using much larger datasets than have been used in previous studies of entrepreneurial emotion. We found that entrepreneurship can lead the entrepreneur to experience more positive general emotions relative to non-entrepreneurs, but to experience less positive emotions directed towards business subjects. Social entrepreneurship can lead to even more positive general emotions than other forms of entrepreneurship, although serial entrepreneurship can lead to less positive general emotions than other forms of entrepreneurship.

- We present the first quantitative study on the entire Twittersphere, that examines the long term evolution of the Twitter network. We observe a network that gets denser through the years, with the number of edges between the users in 2015 being almost double than 2009. We highlight a "rich-get-richer" phenomenon, since the increased number of edges is mainly directed towards the most popular users. Despite the increased number of edges, network connectivity seems to be decreasing. The Largest Strongly Connected component of the network decreases by 20%, in number of nodes, showing that the connections not only increase in total but are also redirected. In the 2009 most of the popular users where popular in both followers and PageRank classification. Our study reveals a decoupling of the two methods, where most popular users through PageRank are not necessarily the ones with the highest in-degree. We identify the reasoning behind users who left the Twittersphere and correlate it with their position in the graph. Our analysis suggests that users who have been banned from Twitter have different degree distributions than others, while the participation in the largest Strongly Connected Component of users who intentionally left the network is by 10% higher than the rest. Furthermore, PageRank classification suggests that several users maintained highly ranked positions before their disappearance.

To this end, the rest of the document is organized as follows: Chapter 2 outlines the related works. Chapter 3 presents a framework for efficient crowd crawling of OSN, while Chapter 4 presents an effective methodology for identifying a user's key locations, namely her home and work places, with real-life applications. Chapter 5 showcase our work on analyzing entrepreneurs sentiment in OSN. Our awarded work on revisiting, analyzing and comparing the complete Twitter graph is presented

in Chapter 6 and, finally, Chapter 7 concludes this thesis.

Chapter 2

Background - Related Work

The chapter is divided in two parts: we first categorize and present the required background through research studies in the field of Online Social Networks analysis, and we then going in-depth by presenting the state-of-the-art related work and scientific methods in the main problem domains that this thesis addresses.

2.1 Background

In this section we overview interesting research works that study the structure and the content shared in OSN platforms. Social networks, referring either to physical or online communities, could be used as a rich source of knowledge extraction. Research communities have shown special interest in the past years in studying a variety of social network communities in order to extract any kind of valuable knowledge according to their studying interest. The field of analyzing and understanding these aspects is of high interest for the research community, and enables a comparison with physical world social networks. Furthermore, knowledge about the users behavior in such a large scale gives the ability of creating models and simulations of social interactions. We divide this chapter in two sections: *i*) *Topology characteristics*: where works which describe topology characteristics of the networks are presented and, *ii*) *Content and Users characteristics*: where we include works related with the users behavior and the type of the content that is generated through the use of these platforms.

2.1.1 Terminology

Here we briefly introduce and explain the terms that we use in the rest of the chapter, related with Topology, Content and Users characteristics that will be described:

Social graph: OSN are modeled using traditional graph theory fundamental concepts. In a graph $G = (V, E)$, each node $v \in V$ represent a user of the respective social network while each edge $e \in E$ represents *a direct connection between 2 nodes* $\{v_s \rightarrow v_d\}$. This connection could be directed (e.g. Twitter following relationship) or bidirectional (e.g. Facebook friendship relationship).

Node Degree and Average Node Degree: Is the number of edges connected to the node. In directed networks in-degree is the number of edges who have the node as *destination*, while out-degree is the number of edges who have the node as *source*. The degree is the sum of in and out-degree edges. Average node degree represents the sum of the nodes' degrees k of a graph divided by their quantity N:

$$K = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{2.1}$$

Shortest path and Average shortest path length: Shortest path between two vertices (or nodes) in a graph is the one such that the sum of the weights of its constituent edges is minimized. In social networks of non-weighted edges, shortest path is the smallest route (regarding number of hops) between two nodes. Average shortest path length is the sum of the shortest path lengths between the nodes divided by the total number of nodes.

Multilayer networks: Multilayer networks are combinations of underlying networks of the same node in different overlay topologies. Such networks can be constructed from information retrieved from Online Social Networks, open-data, etc. For example: *i*) Colleagues network and *ii*) Skills network, where nodes represent users and edges represent *i*) colleague relationship, where connected nodes are colleagues and *ii*) skills relationship, where connected nodes share the same skill. In both networks, nodes v represent the same information (users), while each edge $e \in E$ represents *a relation between 2 nodes* $\{v_s \rightarrow v_d\}$. Such relation could be friendship, interaction relationship, sharing same skills/interests relationship.

Diameter and Effective Diameter: The diameter is the length of the longest shortest path in a graph. Effective diameter is the 90th percentile diameter [62]

Small-World properties: Milgram [72] perform an experiment in order to measure the number of hops that are required for a message to travel between a random pair of people in the United States. Practically, the experiment measures the length of the average path of the offline social network. The results of the experiment show that the average path length was about six hops.

2.1.2 Topology

OSNs maintain enormous directories of users which are increasing rapidly on a daily basis. These users are practically different nodes in such networks, while the relationships between them (friendship, follow etc.) represent the edges in these large scale graphs. The analysis of OSN topological properties is of high interest for the research community, as it reveals in-depth information about the construction dynamics of the graph and also knowledge concerning people communities and societal patterns. Researchers across fields, such as Social Scientists, have the capacity to study and validate their different theories and models in scale, as the digitalized format of the datasets makes them easily accessible.

In recent years, the study of the topologies of the different OSN platforms have gain a lot of attention. Several platforms such as Cyworld, Facebook, MySpace, Orkut, Twitter, Youtube, Flickr, Foursquare, LiveJour and MSN messenger have been widely studied. Furthermore, intra-organizations enterprise social networking platforms, have also attract the attention of researchers, as they reveal information about the interactions among people organized groups, such as employees of a company [13].

It's a Small World

Networks constructed by people and their relationships in the physical world have been extensively studied. During the last years scientists have performed different experiments where they form hypothesis regarding the relationship between people and the overall structure of the underlying networks, with the most famous being the Milgram's experiment which lead to the famous small-world phenomenon [72].

The nature of Online Social Networking platforms enables Leskovec et al. [61] to study the Milgram's theory, using data generated by users of MSN messenger. In specific, they study the mean distance between users who interact through this platform. Their results show that the average degree of separation is 6 intermediaries

and people who share similar physical world attributes, such as age and locations, tend to create connections and maintain a more frequent communication between each other than with users whom their characteristics differ. Furthermore Leskovec, Kleinberg and Faloutsos [62] perform a study on the evolution of 4 real-graphs (ArXiv citation, U.S Patents citation, Autonomous Systems-AS and affiliation graphs) aiming in identifying the growth patterns of such networks. The assumptions that the average node degree remains constant and the diameter is slowly growing over time are examined. Surprisingly, they observe that the aforementioned assumptions and common-truths do not hold. Contrariwise, they show that the networks are becoming denser over time, with the average degree increasing; these results show that the number of edges grows superlinearly in the number of nodes. Furthermore, they show that the effective diameter of the networks is decreasing as the network grows over time.

Mislove et al. [74] collected a large dataset of users from four popular social networks who belong to the large Weakly Connected Component[1]of each graph. By analyzing such data they manage to confirm the ”small world” properties of the online social network. Moreover, the networks degree distribution seems to be in exponential form and the number of incoming links to each node (in-degree) is usually equivalent to the number of outgoing links (out-degree) of that node.

Ugander et al. [99], study the communities in Facebook graph and the structure of the connected components that exists in the network. Their results show that the social network is nearly fully connected as more than 99% of the total users directory are part of the same large connected component. Furthermore, they proceeded in analyzing communities constructed by friends and showed that, despite the fact that Facebook graph as a whole is clearly sparse, the sub-graphs representing neighborhoods of users have a surprisingly dense structure. Also, their results suggests that similarities in nationality and age are strong factors in the generation of friendship connection. Through their study they manage to confirm the six-degrees of separation phenomenon in Facebook, the largest social network regarding active users. A latest study from Facebook[2]shows that the average degree of separation in the net-

[1]A weakly connected component in a directed graph is a set of nodes where each node in the set has a path to every other node in the set if all links are viewed as undirected.

[2]Three and a half degrees of separation, https://research.facebook.com/blog/three-and-a-half-degrees-of-separation

work gets smaller and reaches an average value of 3.57 intermediaries while within the US, people are connected to each other by an average of 3.46 nodes.

2.1.3 Content and Users

In OSN platforms users tend to construct their profiles and share different type of information, according to content being shared in the respective platform. During recent years, the challenge of understanding the users' behavior in the variety of platforms has been widely studied.

McAndrew et al. [70] study how people use Facebook and the factors that influence them, by analyzing an international sample of more than 1000 users. Their results show that females engaged in far more Facebook activity than did males. In addition the former spent more time on the platform and have more friends. Males are less interested in the relationship status of others and expend less energy than women in using profile photographs as a tool for impression management and in studying the photographs of other people. Moreover, males are more interested in how many friends their Facebook friends have but less likely to read the educational and career related information than females. Regarding the age factor, older people spent less time on Facebook, they have fewer friends, and generally use less platform's functionalities than younger people did.

Behaviour of users in different OSN platforms has been studied by Maruf et al. [68]. In their work they have collected a set of 102 users, for whom they retrieve their Disqus and Twitter profiles and perform a linguistic based analysis using the LIWC tool[3]. They perform wide-spread investigation on how personality traits, human interest and sentiment differ in the use of these two OSN platforms. Their findings show that characteristics such as openness, neuroticism and conscientiousness are strong in a user's Disqus comments whereas extraversion is strong in tweets. Furthermore, 80% of the users have more than 65% of their discussed topics differ in two networks, while the common topics discussed in their posts are less than 20%.

Krishnamurthy et al. [53] aimed to characterize Twitter users by analyzing a dataset contained 100,000 of them. They proceed to the identification of distinct classes of users and their behaviors along with geographic growth patterns. They group users based on their ego networks structure: *broadcasters*, who are the users

[3]Linguistic inquiry and word count: LIWC, http://liwc.wpengine.com/

who follow much less users that their followers, *acquaintances* who tend to exhibit reciprocity in their relationships, and *miscreants or evangelists* who follow many more users than their followers and are usually categorized as spammers or stalkers.

Benevenuto et al. [6] study the behavior of the user from a different angle. Their study is based on detailed click-stream data, collected over a 12-day period, summarizing HTTP sessions of 37,024 users who accessed four popular social networks: Orkut, MySpace, Hi5, and LinkedIn. Their results show that 92% of user's active time in OSN services is spend without posting any content, but just for browsing other user's profiles. Moreover, they observe a very low degree of interaction as the average user interacted with 3.2 friends in total over the 12-day period, and interacted visibly with only 0.2 friends. This fact has been also observed by Wilson et al. [105] who showed that in Facebook social network nearly 60% of users exhibit no interaction at all over an entire year.

The behavior of the users in OSN platforms has also been studied by Morris et al. [75]. In this work they constructed a survey study in order to infer the type of questions that users ask in these platforms and the answers that he gets. They use a sample of 624 participants who gave information on what motivates them to respond to questions seen in their friends' status messages. This work concluded that people chose to post questions to their networks because they knew their networks formed a specific audience that they believed to be particularly knowledgeable about a topic. The most popular categories of questions and answers in these platforms are Technology, Entertainment, Home & Family.

Special categories: Students and Employees According to Cheung [16], Facebook is currently the most popular online social networking site among students. Hew [43] examines the usage of Facebook from students perspective - the main education actors - through a comprehensive survey. The results show that student's use Facebook to communicate with friends and maintain their relationships, meet new people and express their selves by sharing thoughts, ideas and information. Additionally, the fact that Facebook is a popular platform and a media that can be used to help an individual to gain popularity is also a motivation for students. Students also use Facebook for educational purposes, such as sharing assignments with friends or access information regarding different courses. An average student has a community of 150 to 350 friends who belong to a similar age group as them.

Bozzon et al. [11] study the behavior of employees in social networking platforms with an emphasis on LinkedIn professional social network. Specifically, they study to what extend enterprise organization information is implicitly revealed by users in their LinkedIn profile and if key organization employees can be identified by observing publicly available information. Moreover, they summarize the factors, related to the employer and the operational organization, that influence the social reach of an employee in the social networks. From their results we can see that the majority of users maintain a profile in different platforms, where they share different types of information. Employees tend to use a different platform for their professional interactions, where they interact with their career related connections, e.g colleagues. Furthermore, after analyzing the employees' profiles in LinkedIn, they show that the majority of IBMers reveal their internal organization job role in LinkedIn. According to the terms that employees use there is a clear linguistic distinction in the job role description of managers and non-managers, thus suggesting an easy identification of important coordination figures within the organization. Regarding the ego-networks, managers tend to connect more with employees higher in the organization's hierarchy.

2.1.4 Knowledge extraction from OSN

Online Social Networks have gained an increasing popularity during the last years, with an enormous stream of information being available for consumption. Due to their user-friendly and usability-oriented interfaces, these platforms attract users who belong in different demographic groups regarding their gender, age, education, job, geographic location etc. Users of such platforms construct their social graphs, share content and interact with other users. For these reasons, OSN platforms are an important source for data retrieval and analysis to drive into conclusions or validate theories in a variety of fields such as social science, market analysis, transportation. For example, for social science researchers OSNs are valuable information sources that can be used in human behavior studies [89], while in the field of computer networks can be used in the design and deployment of new architectures [45].

Real-life Insights

In OSN users are able to easily share different types of information with their communities. Such information includes descriptions about various incidents that happen at a certain time and location, along with opinions and judgements. These facts generate a new challenging and attractive field of study for the data analysts: Understanding the real-world incidents through these enormous streams of data.

Solving Social Problems Twitter gain most of the attention due to its API, which enables the researchers to collect the publicly available information. Sakaki et al. [93] use Twitter users from Japan as social sensors, in order to identify earthquakes in real-time and broadcast warnings. They are able to detect 96% of earthquakes that were classified as strong, based on the Japan Meteorological Agency scale. Authors prepare the training data and devise a classifier using a support vector machine based on features such as keywords in a tweet, the number of words, and the context of target-event words. Their system is able to notify the users by email within 20 to 60 seconds, which is by far faster than the rapid broadcast of announcements of Japan Meteorological Agency, which are widely broadcast in the media.

Twitcident is another early-warning system based on Twitter data, proposed by Abel et al. [1] that aims in detecting incidents, which require the assistance of police, fire department or other public emergency services to take an action. Their system collects the Twitter stream and creates incidents profiles, where the locations and the type are stored. Then they aggregate and semantically enrich the different collected data. Another step for detecting entities such as persons, locations or organizations that are mentioned in tweets is required, and is made possible with the use of existing platforms, such as DBpedia spotlight and Alchemy API. Berlingerio et al. [7] propose a similar system, which identifies and characterize public safety related incidents from social media, and enriches the situational awareness that law enforcement entities have on potentially-unreported activities happening in a city. The system uses Tweets as social media reports and propose a spatio-temporal clustering algorithm that is able to identify and characterize relevant incidents. Specifically, SaferCity retrieves geo-tagged tweets from Twitter and geo-located photos from Flickr and Panoramio, along with information about public events, reported incidents and citizens complaints from law enforcement authorities. A semantic labeler is then

responsible to provide enrichment of the words contained in the social data content using an offline vocabulary, called IPSV. In data sources like Twitter, where the length of the content is limited, the information extraction task is hard, as posts are usually short, noisy, contain ungrammatical text and provide very limited context in the words they contain.

Paul and Dredze [87] examine the applicability of Twitter data in the broader public health research field. In their work they apply the Ailment Topic Aspect Model (ATAM) to create structured disease information (that are used for public health metrics) from Tweets that they have collected. They show that the system has the ability to mine public health information, based on certain keywords (diseases and symptoms) both over time and geographic locations. Additionally, they performed an experiment on allergy season ranges over the different dates by region and observe that they are able to identify several known patterns of seasonal allergies. Their results show that Twitter could be a potential resource for health officials and researchers and can replace current expensive and time-consuming methods.

Trends and Events Detection in OSN Detecting popular topics is another field in online social networking research. Guille et al. [38] suggest that their are mainly two ways to detect such patterns, by analyzing *i*) text regarding terms frequency or *ii*) graph by social interaction frequency. Shamma et al. [94] use a normalized term frequency to demonstrate how an effective table of contents can be extracted by finding localized "peaky topics". According to authors "peaky topics" show highly localized momentary terms of interest. They also investigate persistent conversational topics that show less salient terms which sustain for a longer duration, called "persistent conversations". To mine text across these two metrics, authors introduce a simple term scoring approach, similar to the well known tf-idf, which takes as input all the terms tweeted over a given time frame. Their approach expects that "moments of interest" will have terms associated with them, which appear frequently in the temporal vicinity of the event and relatively infrequent at the rest time-frames.

Furthermore, Weng and Lee [102] aim in identifying trending topics by representing the frequency that a word or phrase appear as wave. Their system first generates signals which captures the bursts in the frequency of words usage. Then it filters out trivial words, by analyzing their signals. By clustering the remaining signals using modularity-based partitioning it aims in identifying a trending topic.

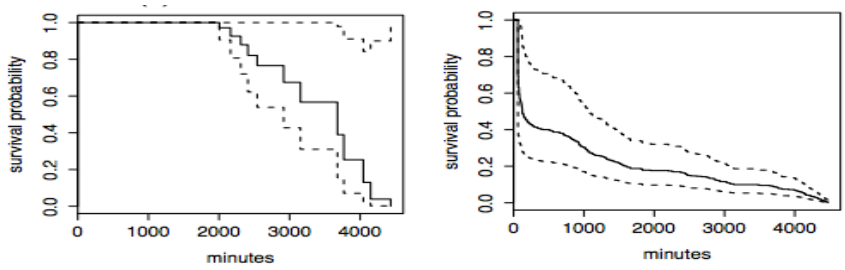
Each cluster is a trend and its popularity is measured based on the number of the words and the cross correlation among the words relating to it. Becker et al. [5] investigate approaches for analyzing the stream of Twitter messages to distinguish between messages about real-world events and non-event messages. The challenge that they aim to solve is that in Twitter exist a number of Twitter-centric trending topics, that are meaningful only in the intra-network community of Twitter. However, these trends usually show similar temporal distribution characteristics with real-world events. This work focuses on online identification of real-world event content, based on a classifier which is able to distinguish the real events related topics and Twitter-centric. The proposed methodology uses an online clustering technique which groups together similar tweets based on their terminology and computes revealing features for each cluster to help determine which clusters correspond to events.

People tend to share content related with several places during their visits or afterwards. This action motivate Choudhury et al. [23] to build a framework for automated generation of travel itineraries using data retrieved from OSN platforms. They develop a two step approach, where the first part takes as input a city and retrieves the photo stream of individual users in the city from Flickr. Then they aggregate all the photos collection in their corresponding POI. Finally, the itineraries are automatically generated from the POIs, based on their popularity and subject to user's time and destination constraints. The quality of travel itineraries constructed by the proposed system was evaluated through a user study conducted using Amazon Mechanical Turk framework.

2.1.5 Information Dissemination and Influence

The content that is being shared in Online Social networking platforms has a lifecycle: it gets born at the time that it is generated and posted, grows becomes popular and survives and eventually stops being reproduced and dies. Moreover, as in real-life in the online social network there exists high and less influential users. Therefore, information dissemination is largely defined and constrained by the aforementioned factors.

Lin et al. [65] examine the growth, survival and context of 256 newly introduced Twitter hashtags during the 2012 U.S. presidential debates. According to them,



(a) (b)

Figure 2.1: Estimated survival time for the two classes. The dash lines show a point-wise 95% confidence envelope around the survival function. (courtesy of [65])

hashtags usually reflect at a topic their emergence is 'happenstance'. They construct the hashtags trajectories, and divide them in two classes: *i*) *winners*, which are used for longer periods and *ii*) *also-runs*, which are the rest. Their results suggest that the number of followers a user has in not a significant predictor of hashtag's longevity. Furthermore, more replies on a hashtag increases the longevity of 'winners' but does not have any effect on 'also runs'. Furthermore, 50% of the 'winner' hashtags lives for 2.5 days, while the same fraction of 'also-runs' dies within 2 hours, as plotted in figure 2.1.

In their extensive study on Twitter, Kwak et al. [59] examine the active periods of trends. A trend is defined as inactive if there is no tweet on the corresponding topic for 24 hours. They observe that 73% topics have a single active period, 15% of topics have 2 active periods and 5% have 3, while very few have more than 3 active periods. Furthermore, the majority of the active periods have a duration of a week or less. 31% of periods are 1 day long, and only 7% of periods are longer than 10 days. Furthermore, they build and analyze the retweets graph and observe that no matter how many followers a user has, the tweet is likely to reach a certain number of audience, once the user's tweet starts spreading via retweets. Figure 2.2 plots the time lag from a tweet to its retweet. Half of retweeting occurs within an hour, 75% during the same day, while about 10% of retweets take place a month later.

The time that an information item is published on an OSN platform is important for the audience that will eventually access it, and therefore it affects its life-cycle. Spasojevic et al. [95] state that the probability of receiving reactions to a posted message is affected by different factors like location, visibility of the message, daily and weekly interaction patterns. In their work, they study how information is

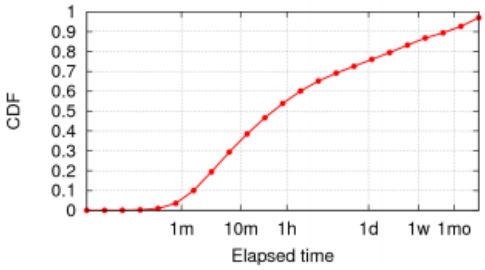


Figure 2.2: Time lag between a retweet and the original tweet (courtesy of [59])

disseminated and its life-cycle in Facebook and Twitter according to their broadcast times and audience. Their results show that a majority of reactions occur within the first 2 hours of posting times on most networks, while audience behavior is significantly different on different networks, with Twitter having larger reaction volumes in shorter time windows as compared to Facebook.

Naveed et al. [81] study what causes a Tweet to be retweeted, with focus on the content of a tweet. They observe that a tweet is likely to be retweeted when it related with a general public topic instead of a personal one. For example when a tweet is about Christmas or social media is more likely to be retweeted than one which addresses to another user directly. They justify this fact based on the nature of Twitter, as it is characterized more as a news and announcement channel rather than personal communication platform; similar observations where also presented by Kwak et al. [59].

Starbird and Palen [96] examine the microblogging information diffusion activity during the 2011 Egyptian political uprisings. Specifically, they study the retweets as means to understand broader Twitter behavior around these protests, and to demonstrate how remote individuals participated in Egypt's 2011 revolution through OSN activities. In their findings they show that 30% of the 1000 top retweeted Twitterers in the usage of popular hashtags related to the protests, were physically located in Cairo during the events. After an analysis of the content, they reveal that tweets contained information about meeting times, injuries, supplies needed etc. Thus, they conclude that revolutionaries were disseminating information and coordinate their actions through Twitter.

2.2 Related Work

In this section we present the related studies on the main problem domains, which this thesis focuses. We first present the work done in the area of Data Collection from OSN. We then turn our attention in the identification of locations using OSN data. Next, we present the studies in the domain of sentiment analysis and conclude the section by presenting the state-of-the-art results in OSN evolution.

2.2.1 Data Collection in OSN

Cho et al. [18] study the design of an effective web crawler. They present several problems in crawling procedure created by the rapid increment on the size of the Web. They propose multiple architectures for parallel distributed crawling framework and identify the challenges in the field of crawling the Web, similarly with [24]. Several challenges are also identified in the design and implementation of an effective large scale dataset collection framework, as the increasing quantity of information that is published in OSN platforms introduce relevant problems. The majority of these platforms maintain monitoring services to control the data throughput, introducing several additional challenges to the parallel data collection campaigns.

A major problem in a data collection campaign is the one of requests rate limiting policies of OSN providers. In the recent years major OSN platforms have used IP-based policies, which restrict a single machine to perform a certain number of requests [74]. The solution on addressing this challenge was straight-forward: a distributed data collection procedure was able to effectively overcome this limitation. Ding et al. [25] present the different categories of challenges for building a crowd crawling system, highlighting the resource diversity of the different parts, the different rate limiting policies from OSN providers, and the data fidelity. They propose a framework of crowd crawling, where a team of multiple research groups share resources in order to efficiently collaborate in a data crawling procedure. Their prototype is implemented over Planetlab, from which they take advantage of the availability of multiple nodes with different IP addresses.

Coalmine [103] is a social network data-mining system, which implements its own mechanism for collecting the data from Twitter and is able to retrieve data using

the official API. Its overall architecture is based on distributed principles, where multiple IP addresses are used. Gjoka et al. [36] propose another similar framework for large scale dataset collection from Facebook. They design and implement a distributed tool which is able to overcome IP-based limitations and collect a large sample. SMIDGen [69] also aims in the collection and extraction of large-scale datasets from OSN. They present a model which is able to efficiently collection content from Youtube platform, following its IP-based policy.

However, OSN platforms, such as Twitter, have changed the IP-based policy to Application-based. The latter restricts a single application from performing a large number of requests. Thus, this update makes a large scale dataset collection procedure more complicated, as the distributed design in the proposed fashion is not functional. With our work we propose a framework which is able to overcome the newly introduced challenges in the field and perform a large scale data collection campaign in the most efficient way. Furthermore, a basic low-resource demanding configuration of the proposed solution enables the collection of more than 2M complete user information in one day, while state-of-the-art requires a much more resource demanding configuration to achieve similar performance.

2.2.2 Identifying locations in OSN

Geo-tagged based approaches Georgiev et al. [35] aim to study users' geographic activity patterns using data retrieved from Four-square. In their study they aimed at estimating user home locations and investigate the influence in events participation. The home location of the user is of high importance for their study and they assume that a user's home place is his most popular place as estimated by the number of Foursquare check-ins.

Jourdak et al. [47] investigate the influence of home location to user mobility patterns, also by marking the most frequently visited location as home place. This approach is probably the easiest and fastest in inferring home places, however it lacks on accuracy and granularity. As it used, it could affect the results of such study and thus its contribution to research community.

A similar approach, proposed by Cho et al. [17], divides the geographical space in 25 by 25Km cells and define the home location of a user as the average position of him in the cell with the most check-ins.

Hawelka et al. [40] investigate global mobility patterns with the use of Twitter. For that purpose they need to estimate users' residence country. They mark as country of residence the country where the user published most of her tweets.

Sadilek et al. [92] described a location estimation method that can infer the most likely location of people for a given time period from the geo-location information of their friends for that time period. They have implemented both a supervised and unsupervised version of their algorithm. In their supervised approach, previously visited locations of users are also given to the prediction algorithm in addition to their friends' locations. In the unsupervised approach, such information (user's previous visited locations) is not given to the algorithm. For the unsupervised approach, they have demonstrated that when a person has at least 2 geo-active friends for whom geo-information of tweets are available, the location of the person can be predicted at the neighborhood level (e.g., a foursquare venue) with 47% accuracy using their algorithm and when 9 geo-active friends' information is available, location can be predicted with 57% accuracy. These accuracies are higher with supervised approach (77% with 2 friend's information and 84% for 9 friend's information). However their approach is dependent on one's geo-active friends (who post messages with geo-location at-least 100 times a month), and the availability of geolocation information for such friends for a given period. In addition, their location prediction algorithm also assumes that a set of locations (e.g., foursquare venues) are frequently visited by users. These assumptions may not be valid for many users who do not have such friends or do not frequently visit such popular locations.

All the above use the average, median or most popular coordinate to estimate the location of the user, without considering her different daily habits. A simple drawback of these approaches is that a regularly visited place, like a cafeteria or the cinema, would have a significant impact on the identified location. Based on this we proposed a method that takes into account the different hours of the day that the user will most likely reside in a Key location, thus eliminating, to a high degree, the influence of user's hot spots.

Content based approaches Mahmud et al. [67] propose an approach that based on a location dictionary for places all over the United States, manages to infer 57% of users Home Location using their Tweets at city-level granularity. They present a hierarchical classification approach which narrows down the granularity from

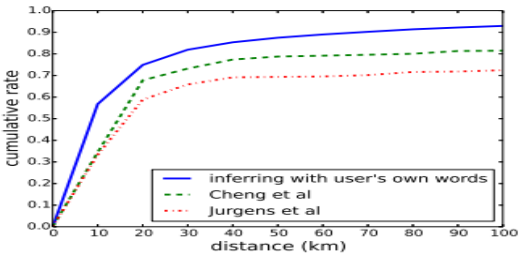


Figure 2.3: Performance of methods. Ryoo et al. method [91] outperforms the others in all distance sections. (courtesy of [91])

timezone, state or region and then city.

Ryoo and Moon [91] propose a content-based approach that aims in identifying Twitter users' home place in a granularity of 10Km, using the tweet textual contexts. They use a probabilistic model to assign location data to popular words in Twitter and then use the popularity of these words to identify the location of the users that tweet them. Their approach manages to identify up to 57% of the users in a 10Km radius and outperforms the compared methods.

Li et al. [63] present an effective location identification approach based on information collected from multiple microblogs, combined and utilized in order to identify the top-k candidate locations of a user. They show an accuracy similar to that of Ryoo and Moon. Chong et al. [19] aim in identifying specific venues, based on the content of the tweets. They formulate fine-grained geolocation as a ranking problem, whereby given a tweet should rank and propose candidate venues based on other locations that users has tweeted from.

All the above approaches use the aforementioned geo-tagged methods as ground truth for the users' home location. As mentioned before these approaches lack in terms of accuracy and thus are not able to provide valid ground truth information. Our approach significantly improves the accuracy of those techniques, thus can be additionally used to provide the ground truth information needed from context-based approaches.

Applications Noulas et al. [84] study the check-in dynamics of a large number of Foursquare users for a period of 100 days. They conclude that location information reveals meaningful patterns in both mobility and urban spaces. In specific, they show that such information unfolds transitions between different types of places and an analysis on the geo-temporal rhythms of user check-ins is able to provide us

with insights about a users activities.

Cho et al. [17], after identifying a users home location, study the basic laws govern human motion and dynamics using Growalla, Brightkite and cell phone location data. From the results they conclude that social relationships can explain 10%-30% of human movements. Zhang et al. [110] propose and implement a novel architecture which is able to explore human mobility using multi-source data. Their system infers users locations from various resources such as cellular networks and transit networks. Based on their evaluation, they are able to infer mobility patterns with an average accuracy of 75%.

Cici et al. [21] identify a user's home locations, using most popular place as home, and then study the potential of ride-sharing based on mobility patterns extracted from cell phone and social networks data. They propose an algorithm that divides different users in groups, based on the similarities in their mobility patterns. Their results show a decrease of 31% in car usage, when users are willing to ride with friends of friends.

2.2.3 Sentiment in Online Social Media

Online Social Networking and media platforms are popular in the field of sentiment analysis, mostly because they provide access to digital content that has been shared by the user. Milani et al. [71] use Twitter data to propose an approach for the analysis of user interest based on tweets, which can be used in the design of user recommendation systems. Their proposed approach is based on the combination of sentiment extraction and classification analysis of tweet to extract the topic of interest. The topic extraction phase uses a method based on semantic distance in the WordNet taxonomy. Their proposed algorithm has been tested on real tweets generated by 1,000 users, and their results confirm the suitability of the approach combining sentiment and categorization for the topic of interest extraction.

Qingqing et al. [88] apply sentiment analysis aiming to extract the dietary preferences of the user, which currently is a costly procedure mainly based on questionnaires. For their study they use microblogs from weibo.com, to detect dietary preferences of social media users in China via sentiment analysis. Sentiment polarities of the aspects and dishes are identified by sentiment classification. Their results on 3,975,800 microblogs suggest that social media users in China are not satisfied

with their overall dietary, while experimental results show that semantic information is useful in extracting dietary aspects.

Gu et al. [37] study the correlation between the sentiment and information spread in news media. They focus in interactive spiral of online news and examine the relationship between title sentiment and users' different-stage reactions, such as reading, commenting, like/dislike voting and forwarding. Their results suggest that despite the fact that negative titles attract users to read, they negatively influence the user in sharing, as it decreases the number of forwards.

SenticRank [106] is a novel generic framework which incorporate various sentiment information to various sentiment-based information for personalized search by user and resource profiles. The aim of this framework is to obtain sentiment-based personalized ranking in folksonomy [4] and address the problem of the personalized tag-based search in collaborative tagging systems.

Another interesting area with sentiment analysis applications is online shopping. Kaur et al. [52] propose a model for assessing the quality of a product based on the reviews' sentiment. Based on the sentiment analysis they generate a report which shows positive and negative points about the specific product. The proposed model has been evaluated based on the performance parameters of the precision, recall and polarity-based accuracy assessment and results verify that sentiment analysis is a useful tool in the area of online shopping.

Nguyen et al. [83] propose a model for stock price movement prediction using the sentiment from social media. Their approach incorporates the sentiments of the specific topics of the company into the stock prediction model, using data from Yahoo Finance [5]. They evaluate the effectiveness of sentiment analysis in the stock prediction task by performing a large scale experiment. Comparing the accuracy average over 18 stocks in one year transaction, their method achieved 2.07% better performance than the model using historical prices only. Furthermore, when comparing the methods only for the stocks that are difficult to predict, their method achieved 9.83% better accuracy than historical price method, and 3.03% better than human sentiment method. Thus, they showcase the improvements that sentiment

[4]Folksonomy is the process of using digital content tags for categorization or annotation. This process enables user to classify various forms of data (such as websites, pictures, documents), so as to increase the accessibility of the content.

[5]Yahoo Finance: `https://finance.yahoo.com/`

analysis provides for the stock price movement prediction.

Wang et al. [101] propose a latent probabilistic generative model called LSARS to mimic the decision-making process of users' check-in activities both in home-town and out-of-town scenarios by adapting to user interest drift and crowd sentiments, which can learn location-aware and sentiment-aware individual interests from the contents of spatial items and user reviews. Due to the sparsity of user activities in out-of-town regions, LSARS is further designed to incorporate the public preferences learned from local users' check-in behaviors. They evaluate LSARS on spatial item recommendation and target user discovery, using data retrieved form Yelp and Foursquare . Their experiments show that sentiment analysis optimizes the decision-making process.

2.2.4 Online Social Network Evolution

Networks constructed by people and their relationships in the physical world have been extensively studied. During the past years scientists have performed different experiments where they form hypotheses regarding the relationship between people and the overall structure of the underlying networks, with the most famous being the Milgram's experiment which lead to the famous small-world phenomenon [72].

The nature of Online Social Networking platforms enables Leskovec et al. [61] to study the Milgram's theory, using data generated by users of MSN messenger. In specific, they study the mean distance between users who interact through this platform. Their results show that the average degree of separation is 6 intermediaries and people who share similar physical world attributes, such as age and locations, tend to create connections and maintain a more frequent communication between each other than with users whom their characteristics differ.

Furthermore Leskovec, Kleinberg and Faloutsos [62] perform a study on the evolution of 4 real-graphs (ArXiv citation, U.S Patents citation, Autonomous Systems-AS and affiliation graphs) aiming in identifying the growth patterns of such networks. The assumptions that the average node degree remains constant and the diameter is slowly growing over time are examined. Surprisingly, they observe that the aforementioned assumptions and common-truths do not hold. Contrariwise, they show that the networks are becoming denser over time, with the average degree increasing; these results show that the number of edges grows super-linearly in the number

of nodes. Furthermore, they show that the effective diameter of the networks is decreasing as the network grows over time.

Kwak et al. [59], examined the full Twitter graph as it appeared in 2009. Their work summarize the characteristics of Twitter and its power as a new medium of information sharing. Their analysis on the topology and the structural properties of the graph shows an average path length of 4.12, a non power-law follower distribution, a short effective diameter and low reciprocity; all these indications mark a deviation from known characteristics of social networks [82].

With our study we revisit the same sample of users and collect the full information that is available from the Twitter API. We collect a total of 34.6 million user profiles, connected through 2.05 billion relationships. Based on the provided insights and data, we aim in analyzing the Twitter network as is today, and provide a comparison with the snapshot of 2009. We address the different characteristics of the 2009 Twitter network, as it appears to be connected today, and examine the changes in connectivity of the network in general and the users in particular. To the best of our knowledge our work is the first quantitative study on the entire Twittersphere, that examines the long term evolution of the Twitter network.

Chapter 3

Dataset Construction: Retrieving Data for OSN analysis

The content that is generated through the interaction of users in OSN platforms is of high interest for the research community. By analyzing such data a researcher is able to validate her hypothesis and perform high quality studies in the corresponding field. However, the collection of a large-scale dataset is not a trivial task for researchers due to several challenges that are introduced, either by the users with their privacy policies or the resources limitations [9].

3.1 Proposed Framework Design

In this section we present the design of the proposed framework, by introducing the basic components and their core functionalities. Furthermore, we describe the communication between the different components of the system, and how each one of them contributes to the goal of increasing the efficiency in a large scale dataset collection campaign.

OSN platforms, such as Twitter, Facebook and LinkedIn, represent information in similar abstractions. Each user has a unique identifier, usually of type *long*. Similar identifiers are assigned in messages posted by the user, like *Tweets* and *Posts* in Twitter and Facebook respectively. The connections between the users are retrievable as edge lists, which denote either a reciprocal (friend) or direct (follower) connection between two users. The retrieval of this information can be achieved either using OSN provided API or through Web Scraping. However, the terms-of-services of

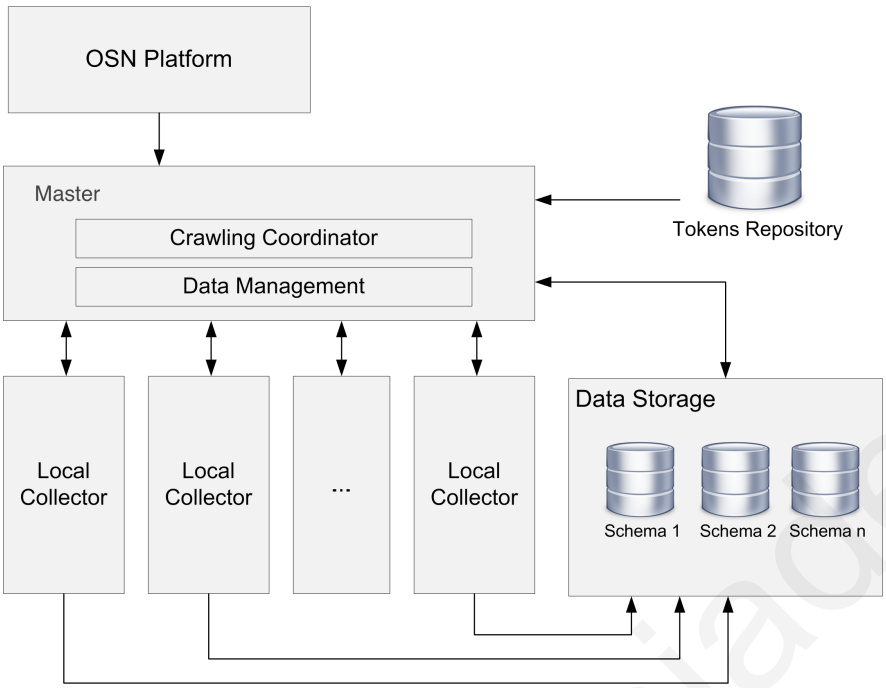


Figure 3.1: General System architecture

popular OSN prohibit the data collection through automatic Web Scraping [1] [2].

A data collection campaign can be either *i*) Resource Specific collection or *ii*) a Real-Time stream collection. The first case provides retrieval services for a specific resource (e.g a user's profile, a tweet or post), while the latter enables the sample collection of real-time information that is being published in the OSN. Our proposed framework provides two different services in order to enable both cases of a data collection campaign, which are 1. Resource Specific Data Collection, and 2. Real-Time Stream Collection.

3.1.1 Crowd Crawling: Building the Tokens Repository

As mentioned in section 2.2.1, a newly introduced challenge in data collection procedures is the update of IP-based limitations to Application-based ones. For a complete presentation of our crowd crawling approach we first describe the traditional data collection procedure, established through the available OSN API. An individual who aims in using the services of an OSN API is required to create an application in the specific platform. In our case, the requests are related with data collection, either of a specific resource or the public OSN stream. By creating the application, the

[1] `https://twitter.com/tos`, Twitter Terms of Service (Last Accessed: October 2016)

[2] `https://www.facebook.com/apps/site_scraping_tos_terms.php`, Facebook Terms of Service (Last Accessed: October 2016)

user agrees with the terms of service, and the OSN is able to monitor the requests executed through it. OSNs API policies restrict a single application of performing a large number of requests. In order to make authorized calls to an OSN API, each application must obtain a group of access *tokens* on behalf of a user, usually through the OAuth 2.0 specification [46]. These tokens are unique for each application and each user. For example, in the case of Twitter API, the platform provides four tokens; two are related with the application while the rest are related with the user who agrees to authorize the application to execute requests on user's behalf. Furthermore, the user has the ability to revoke the generated tokens at any time. This will result to the deactivation of the generated tokens; an action that will restrict from the application to execute API requests on user's behalf.

In IP-based limitations, the OSN API monitors the public IP address of the machine and applies the limitations per IP. Thus, a distributed data collection campaign in different machines, but with the same group of tokens, radically improves the procedure [10,100]. However, after the latest updates, this is not possible as the majority of OSN API monitors the registered applications. As a result, even when an application (that utilizes the same set of tokens) is distributed in several machines with different public IP addresses, the limitations that apply are the same as if it was running on a single instance. Having this in mind, we proceed in a crowd-crawling approach, asking from OSN users to contribute to the data collection procedure by authorizing applications to access the API.

The procedure that we follow is the one suggested by the OSN platform. We first register an application in the OSN API. Then, we develop a service which asks from the users of our ego-networks (followers and followees) to authorize it to execute public data retrieval requests. Having the approval of the user, our service collects the generated tokens and stores them in a *Tokens Repository*. This repository contains a number of tokens that have been generated by OSN users. This crowd crawling procedure increases the number of tokens/resources that can be used during the retrieval process in the proposed system and takes place before the beginning of the data collection campaign.

Having multiple tokens enables us to activate a different set of them in order to avoid reaching the resources request limit; when we hit this limit the group of tokens becomes invalid for a certain amount of time t. We then move to the next group of tokens and execute the number of requests until we hit their limit. We follow this

procedure repeatedly, until the condition $t - current\_time = 0$ is satisfied, as the group of tokens will become active again. Thus, with an n number of tokens we enable the continuous operation of the data collection campaign. This procedure is executed in each *Local Collector* instance, which is described in this section.

3.1.2 Resource Specific Data Collection

The proposed service provides functionalities related to asynchronous resource specific data collection. With this term we denote the procedure where we collect resource specific historical data enabling the retrieval and storage of all the available data of a user, given user's unique ID (UID). The proposed framework is able to crawl OSN platforms with the use of parallel API instances.

As depicted in Figure 3.1, the proposed system uses a Map-Reduce-like approach to overcome this limitation by partitioning tokens into a large number of small instances, greater than the available nodes, with some being replicated for performance objectives. Specifically, the system consists of three main components: (*i*) Master Component, (*ii*) Local Collector Component, (*iii*) Data Storage Component. The Local Collector Components are different instances running on different physical machines. Due to the latest API policies, which remove the IP-based requests limitations, multiple instances could be run on a single machine. However, the policies update to Application-based limitations increases the complexity in data collection process parallelization. The Master component is responsible to monitor and maintain the different Local Collectors, taking into consideration the resources demand and availability. The Master component assigns tasks to Local Collectors based on the provided UID list and the available tokens. Through the tokens and UID balancing, Master component manages to maintain the collector resources based on the demand. For example, if a Local Collector does not need the assigned resources, it returns them to the Master component which in turn assigns the resources to a more demanding Local Collector instance. Each Local Collector instance communicates with the Data Storage Component in order to store the retrieved data. The main task of the Data Storage component is to monitor the storage procedure and is able to perform modifications in the storage functionalities in order to ensure a maximum throughput rate. For every action it provides feedback to Master component and proceeds to modifications if needed (e.g. temporary store data in the file system, if

the database engine is down).

UIDs retrieval: The Master component requires a list of UIDs to initiate the data collection procedure. Such a list can be retrieved from the proposed Real-Time Stream Collection service and/or through OSNs public directories. These directories are indexes to the public profiles of users, maintained by popular OSN platforms, such as Twitter[3] and Facebook[4]. The UIDs are used by the Crawling Coordinator, which initializes and distributes the crawling workload to the different Local Collectors.

Master Component: The Master Component is responsible for the workload distribution and monitoring of the data collection process. It has global knowledge about the system's state and maintains the resources based on the corresponding needs, by obtaining up-to-date crawling information from the different Local Collector instances. This component gets as input the list with UIDs that should be collected and calculates the load needed for each local instance. It then distributes the tasks and the resources based on the calculations. When a Local Collector requires more resources it sends a request to the Master component, which will then check the availability and update Local Collectors resources pool. On the other hand, when a Local Collector has reserved resources and does not need them, it notifies the Master component which in turn retrieves them back and makes them available for other resources. With this procedure, the system is maintained in a state where only the required resources are used, and each Local Collector has the highest available amount of the resources that it requires. In general, Master component has as goal to ensure that each Local Collector runs in full throttle 24/7, addressing API requests limitations.

Data Storage Component: The Data Storage Component is responsible for aggregating the anonymized results that have been collected from the different Local Collectors and store them with the most efficient way. This component is able to retrieve the data from multiple instances and store them in a central database. It maintains the storage queues and performs the necessary actions in order to ensure that it does not act as a bottle-neck. It is able to run real-time analytics and inform the Master component about the current metrics and actions taken. Such actions include the creation of different storage schema when the analytics suggest so, use of compression when running-out of space, store data in files when database engines

[3]https://twitter.com/i/directory/

[4]https://www.facebook.com/directory

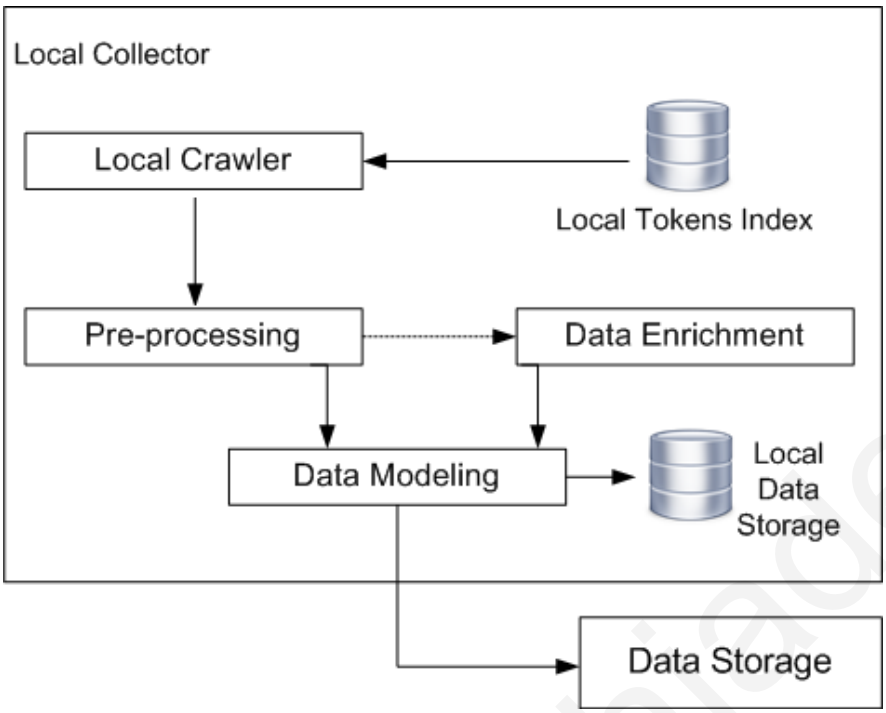


Figure 3.2: Local Collector architecture

fail etc. The component ensures that the retrieved data will be eventually stored in the file system in the most efficient way, at any given time. For the most efficient configuration, Data Storage and Master components should be deployed at the same machine.

Local Collector

In Figure 3.2 we present the architecture of the Local Collector component which is one of the main actors of the system. It is responsible for obtaining a task from Master component and perform the necessary actions in order to complete this task. Additionally, it provides real-time information about the progress to the Master component. Multiple Local Collector instances are distributed and deployed in different machines, increasing the efficiency and the throughput of the data collection system. Here we describe the internal components of the Local Collector and their tasks in the overall data collection campaign.

Local Crawler: In each Local Collector there is a Local Crawler, which is responsible to execute the requests to the OSN platform through the available API. This component uses the local tokens index which has been updated by the Crawling Coordinator of the Master component. Furthermore, it is responsible to maintain the crawling procedure in order to overcome the requests limitations, by requesting or dismissing API tokens. When the Local Crawler hits a request limit, it will automatically inform

the Master component. If the Crawler Coordinator has available tokens, it will update the local index with the new tokens. On the other hand, if the Local Crawler is able to complete the assigned tasks with less tokens, it will inform the Master component and the Crawler Coordinator will dismiss the tokens and update the tokens index.

Pre-processing and Data Enrichment: The retrieved information is first passed through a pre-processing step, where is being cleaned and converted to the require encoding (e.g. convert the non-supported characters to unicode). Furthermore, during the pre-processing step, the retrieved data are being anonymized, according to privacy protection policies[5] and OSN APIs terms-of-service. During the anonymization procedure we replace the user and posts' ids with random numbers. A part of the data is then parsed by the Data Enrichment step, where the collected information is being enriched from external sources (e.g. a Tweet is being parsed to NLTK for sentiment analysis [8], or the post-code of a geo-tagged tweet or post is being identified).

Data Modeling: After, these two steps the resulted data is forwarded to the Data Modeling component, where the final formating applies. Each Local Collector divides the general task in multiple subtasks, that can be executed in parallel on the same instance. For example, when a Local Collector gets the task of collecting 100,000 Twitter users, it is able to execute the crawling procedure in parallel, by running 5 threads which each one collects 20,000 users. Following this procedure the Local Collector is able to take advantage of the local workload division in smaller subtasks and better monitor the crawling procedure. The Local Collector communicates with the Data Storage Component thought a socket. Through this socket, it sends the data that are handled by the Data Storage component and stored at the final step in a database schema. In order to reduce the communication cost, Local Collector is able to temporary store the retrieved data locally and proceed to bulk insertions.

Failure Resistance: Each Local Collector instance maintains a local data storage component, which is activated in cases of failures in the communication with the Data Storage component. Retrieved data are stored in this local component, until the communication is restored and transfered to Data Storage.

[5] http://ec.europa.eu/justice/newsroom/data-protection/news/20150128_en.htm

3.1.3 Real-Time Stream Collection

A main limitation in OSN platforms API is the one of filtering their public stream. Twitter, for example, makes accessible only 1% of the total Twitter stream though the corresponding API [6]. Thus, when a researcher aims at collecting Tweets that are being published from a specific geographical area, e.g. the city of London, she will only be able to retrieve the set that does not exit the threshold of 1% of the total Twitter Stream. Furthermore, Morstatter et al. conclude that the results of using the Twitter Streaming API depend strongly on the coverage and the type of analysis of the study, and highlight the need of methods and frameworks that compensate the biases in these types of API [77].

The proposed framework supports Real-time Stream Collection, a service that is able to overcome these limitations and collect OSN platforms' stream in the most efficient way. This service provides the functionality to retrieve the public stream with 2 different options: (*i*). Given as input a geographic boundary box, (*ii*). Given as input a set of terms. For (*i*) it constantly listens to the stream of the area that lies in a specific boundary box, while for (*ii*) it queries the API for posts which contain the specific terms.

Master Component: Similarly to Resource Specific Data Collection service, the Master component is responsible for monitoring the overall procedure. It takes as input the target file and the Crawling Coordinator distributes the load in the different listeners. For example, in the case of monitoring the stream of a specific location, it takes as input the geographical coordinates of the under investigation area and divides it in a grid. It then distributes the different boundary boxes in a team of Local Collectors, giving them the subtask to collect the stream of a much smaller geographical area. Data Management component is responsible to receive the feedback from the Data Storage and proceed to the necessary actions.

Local Collector: The Local Collector receives a task from the Master component and is responsible to constantly listen OSN stream based on the rules received, using the OSN API. Such rules are a boundary box or a specific set of terms. A Local Collector is also responsible to monitor its resources and ask from the Master component to redistribute the API tokens, and thus the workload, if required. For example, a Local Collector receives a task to listen to the Twitter stream of a part of the city of

[6] https://dev.twitter.com/streaming/overview

London. During rush hours, this area gets crowded, thus the stream exits the limits of the Twitter stream API. At the same time another Local Collector is responsible to receive the stream of a part of Nicosia, Cyprus, which is much less crowded than API thresholds. Both Local Collectors report their monitoring results and request from Master component to redistribute the load. Master's Crawling Coordinator then assigns the resources of the less crowded collector to the crowded one, by creating a sub-grid, while it assigns a nearby Local Collector to the part of the city of Nicosia.

3.2 Evaluation

For the evaluation of the proposed framework we developed a proof-of-concept prototype, following the design requirements presented in section 3.1. We evaluate the proposed framework for both provided functionalities, *Resource Specific Data Collection* and *Real-Time Stream Collection* over several case studies on the Twitter platform. The choice of Twitter for the evaluation was motivated from the fact that the openness of this platform has attracted a large number of research groups to perform analytics and drive into conclusions using data retrieved from its data servers [97]. In this section we present the experimental setting and the results of the experiments. We then discuss our findings and compare with related work.

3.2.1 Properties of Interest

The proposed framework visits a Twitter user's account and collects the following information:
User profile: Each Twitter user is uniquely identified by his UID. In the public profile one can find information about the user's current status (description) and location. Additionally, in a user's profile additional automatic calculated fields can be found, such as tweets, followers and followees, profile creation date and profile image URL.
Tweet: a list of statuses are included in a user's Twitter account[7]. Each *Tweet* entry contains a unique identifier, the UID of the publisher, the text and a set of meta-data. Such meta-data include the timestamp, several flags that denote if the entry is geo-tagged, retweet, reply, favorite, if it has been retweeted and how many times,

[7]We are able to retrieve at most the 3,200 most recent Tweets for each user, due to Twitter API request policy.

the number of mentions and hashtags contained and application that was used to get published. Moreover, for each geo-tagged Tweet, information about the geographical place is included, such as the country, country code, place name, street address, place type and a geographical boundary box. Furthermore, we enrich each geo-tagged Tweet with its corresponding post-code area.

Ego-network: a users ego-network contains a list of followers and followees unique UIDs. The followers list contains edges that are ending on user's profile, while followees list edges that start from the user's profile.

3.2.2 Experimental setting

Our proof-of-concept prototype has been developed in Java. For the data storage component we use MySQL, which is a widely used relational database management system (RDBMS). For the integration between the data collector and storage components we use the JDBC driver.

We deploy a Master component instance on a server with 4-core 2.5GHz processor and 24GB memory. The Master component initiates four different instances of Local Collectors on four different machines, running on 4-core 2GHz processor with 4GB memory each. Additionally, we showcase an experiment on a Raspberry Pi Model B low cost device [8]. This scenario evaluates the execution of parallel instances of Local Collectors, coordinated by one Master component running on the infrastructure of a research institution. The Data Storage component is deployed on the same machine with the Master component, in order to reduce the communication cost between these two actors.

Use Case Scenarios

Resource Specific Data Collection: In the presented crowd crawling case study scenario we use the online social networking platform of Twitter, one of the most widely used platforms in research. In order to generate the UID list we randomly sampled users from the dataset used in [59] and is publicly available. For each UID in this list, Local Collector instances request and store the complete *properties of interest*.

Real-Time Stream Collection: For this evaluation scenario we use the option of collecting the public stream of a boundary box. In order to get better insights

[8] A single-core, low-cost device, running at 700Mhz with 512MB RAM. https://www.raspberrypi.org/products/model-b

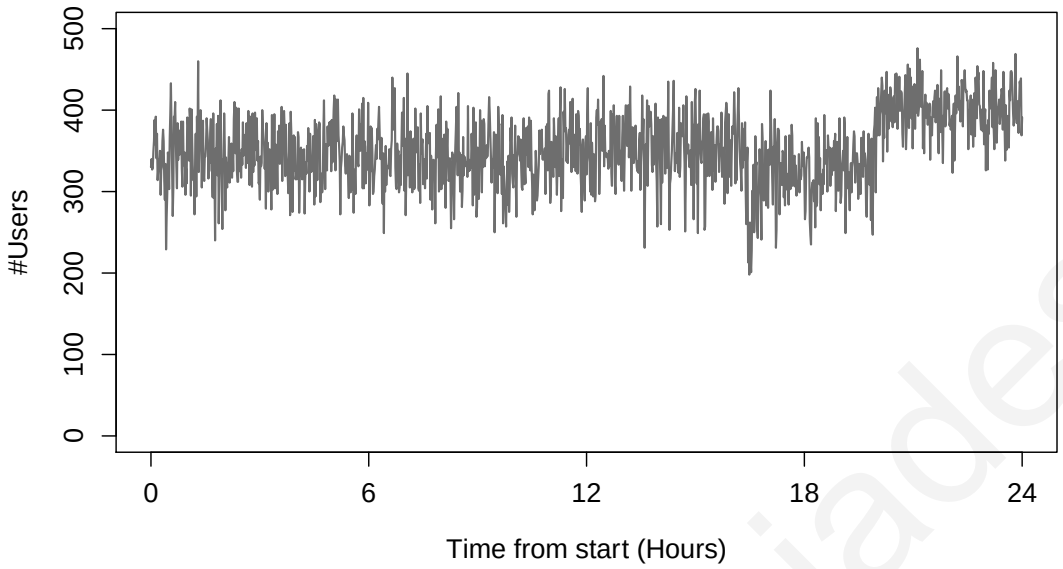


Figure 3.3: Crawling throughput of an average Local Collector component for 24 hours. Each Local Collector is able to retrieve the complete set of Properties of interest for 397.2 users per minute on average.

Users	Followers	Followees	Tweets	Places
2,300,574	1,220,972,850	635,276,364	1,612,766,674	1,040,240

Table 3.1: Number of Users, Followers, Followings, Tweets and Places of geo-tagged Tweets of the resulted dataset.

on the performance we needed a scenario where the threshold of 1% of total Twitter stream will be exit. Thus, we give a boundary box with the complete world map, which indicates that we need to collect the public stream of all the locations. We then execute three different approaches in parallel: We collect the public Twitter Stream of this area using (*i*). *Single* Twitter stream listener using Twitter Stream API (*ii*). *Multiple* instances of Twitter API, listening to the same area, (*iii*). *Real-Time Stream Collection* functionality of the proposed framework. We then compare the results and present the insights.

3.2.3 Results

Crawling Throughput

Resource Specific Data Collection: We perform a distributed crawling procedure,

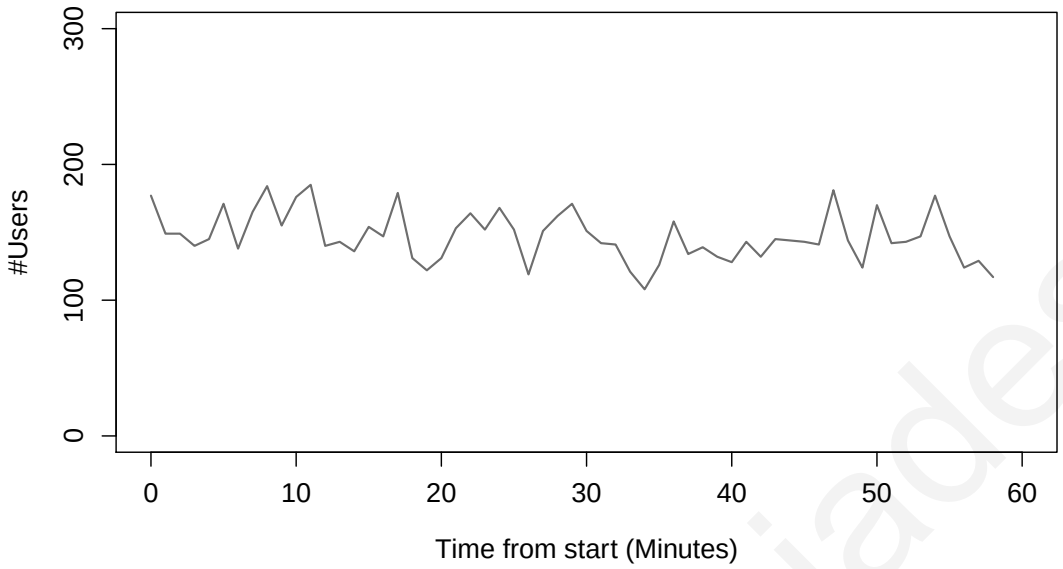


Figure 3.4: Crawling throughput of one Local Collector component, running on a Raspberry PI low cost device. Each Local Collector is able to retrieve the complete set of Properties of Interest for 147.2 users per minute on average.

following the described *Resource Specific Data Collection* methodology, for 24 hours. Figure 3.3 summarizes the throughput rate per minute for an individual Local Collector. Our proof-of-concept was able to obtain more than two million users during this period, having the four Local Collector instances collecting about 575,000 users each, without exceeding 9% of machines' memory usage. An average instance is able to collect and store more than 372 users per minute. During the collection procedure an instance collects the complete *properties of interest* of the requested users, as described above. The resulted dataset, presented on Table 4.1, can be translated in more than 69GB of uncompressed data per Local Collector. Figure 3.4 showcases the performance of an instance running on a Raspberry Pi Model B. As we can see, in one hour of crawling, a Local Collector instance running on such device is able to collect the complete set of *properties of interest* of 8,820 users, a number which is by 3x higher than state-of-the-art [25]. These results show that the intelligent management of resources and tokens radically improves the traditional distributed methodologies.

Real-Time Stream Collection: Figure 3.5 summarizes the throughput rate per minute for the 3 different compared approaches. As we can see, the proposed system is able

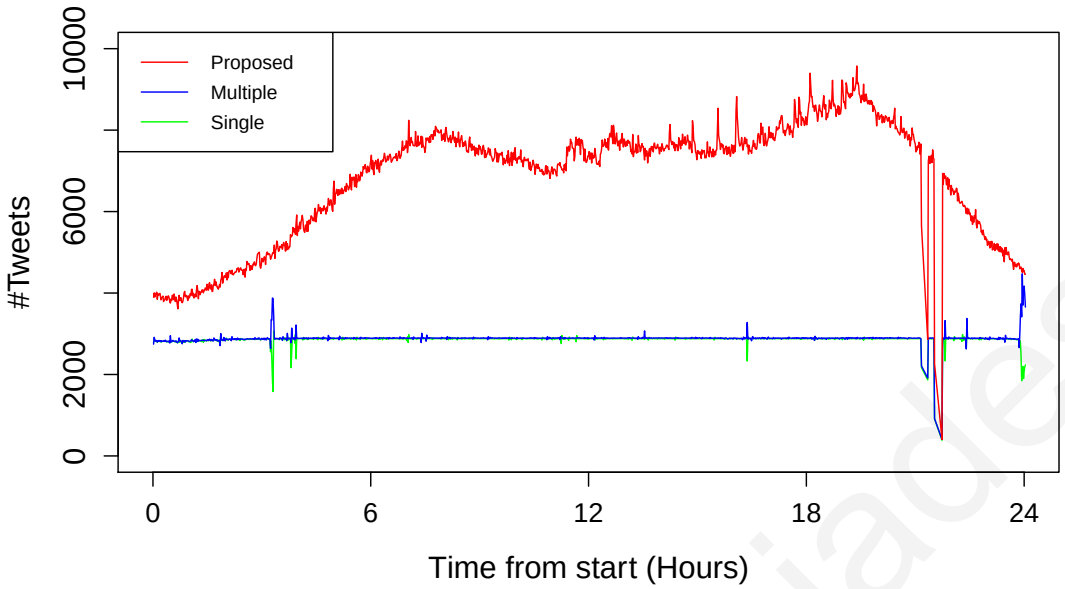


Figure 3.5: Crawling throughput of stream listener, compared with single and multiple instances of Twitter API

to perform a large-scale real-time monitoring campaign with up to 3 times higher throughput than the commonly used approach. The applied procedure on the proposed system resulted to the collection of more than 9M different Tweets, while at the same time Twitter Stream API does not return more than 3M. Furthermore, as we can see from the parallelized procedure, the *Single* and *Multiple Single* instances resulted to similar throughput, having the latter collecting 40K more unique Tweets. Here we should note that the improvement is based on the fact that Twitter limits the access on 1% of the total stream. Thus, in cases of events, the throughput of our platform it will remain about 3x more than the one of the API.

3.2.4 Discussion

General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) is a regulation by which the "European Parliament, the Council of the European Union and the European Commission intend to strengthen and unify data protection for all individuals within the European Union (EU)". [9] This directive also enforces 3rd party data collection tools to follow certain rules and guidelines. The proposed

[9]GDPR: `https://ec.europa.eu/info/law/law-topic/data-protection_en`

framework uses anonymization techniques during the whole procedure, following the directions regarding the data management and storage of GDPR. Furthermore, this framework is only able to collect public data, with respect to users private information.

Summary

The evaluation of our proposed OSN dataset collection framework shows the feasibility of utilizing a number of OSN API Tokens, retrieved through crowdsourcing, to collect a complete and timely dataset, without violating the terms of use of the services. As shown above, through smart utilization of resource, an interested party can collect more data in minimal time, avoiding any bias in the research outcome, created by a long lasting data collection campaign. Additionally, through smart use of resources our framework triples the collection of the real time stream of OSN services. Such an increase can be valuable to both research and commercial application that react based on the real-time census of the active Online Social Network users. In addition to the evaluation performed in this section, the proposed framework was used for the data collection campaigns of [26, 27, 29, 30].

Chapter 4

Inferring Locations from OSN Analysis

The massive adoption of mobile devices that offer Internet connectivity, geo-location capabilities, and continuous access to online social networking services (OSNs) has enabled users to contribute content to OSNs on a continuous basis, from different locations and at different times of the day. Based on this ubiquitous OSN activity, it is now possible to sketch the mobility trajectories of users and to pinpoint their visited locations. However, a large amount of users (34%) do not provide real location information, frequently incorporating fake locations or sarcastic comments that can fool traditional geographic information tools, while only 0.09% had manually entered their location at the precision of an address, a granularity that is higher than city level [41].

In recent years, the automatic mapping of users to their "key" visited locations of interest (e.g., home, work, leisure), based on their online social presence, has been of great interest for the research community [67, 91]. Information about the Key locations that users visit and from which they contribute content to OSNs, has applications in a variety of research fields like understanding user movement [17]; investigating the relation among real-world human activities and interactions, physical spaces, and OSN structure and dynamics [12]; and exploring the challenges to user privacy protection. Moreover, the combined knowledge that can be mined from this information, can be of tremendous help for a diverse number of applications aiming at improving habitats and daily activities in cities, from event identification and recommendation to urban city planning.

4.1 Key Locations Identification

Identification of a user's Key Locations is of high interest for researchers in Online Social Networks analysis fields, who focus either on estimating these locations or enriching their datasets and using them for further analysis. In a large part of the literature, researchers are interested in identifying user home locations. Recent studies present approaches that are focused on estimating a user's Key locations based on geo-tagged OSN activity or/and content that the user publishes in her profile. In several studies they use this information in real-world applications such as studying users' geographic patterns, global mobility patterns, correlation between friendship and mobility. In this section we present studies for both and summarize common methodologies that have been identified.

4.1.1 Problem Formulation

Given as input the geo-tagged Twitter activity T_u of a user u we are interested in the identification of the user's key locations, namely Home and Work locations, denoted as H_u and W_u respectively. The tweet information we are interested in is represented by the vector $< p, t_p >$, where p denotes the geographical coordinates ($< long, lat >$) the user tweeted from at time t_p. The set of all location visited by user u can then be denoted as P_u. Our research then tries to give an answer to the question: *Can we identify a user's u home and work location simply by observing the locations and time the user tweeted from?* In the following sections we introduce a method to answer this research question providing the highest key location identification accuracy and also minimizing the detection radius granularity to as low as possible.

4.2 Dataset

We used a variety of OSNs to collect geo-location information about the users. Regarding workplace location, we introduce a novel method, that combines a variety of OSNs, and datasets.

4.2.1 Home Location

For Home location identification we turn to Twitter and search for users that include geographical information in their tweets. To avoid extensive crawling of the Twitter network we first visit Twitter's live stream for three different geographical areas, namely, the country of Netherlands (March 2014), the city of London, UK and LA county, CA, USA (November 2014). We use the geographical boundaries of these areas and collect geo-tagged tweets within these boundaries. For each of these tweets we collect public information about the user that posted the tweet. This information includes the past tweeting activity of the user, her ego network, followers and followees, and her profile information. As the usage of Twitter Stream API is binded to several limitations [76], we proceeded in further expansion. To expand our dataset we use the users collected from the previous process as *seeders*. For each seeder we randomly crawl users belonging to her ego network and collect the same information. We keep only users that have at least one geo-tagged tweet from the three areas of interest, and add them to the seeders list for further crawling.

Data cleansing: One major concern for any Twitter dataset is to avoid bots, which act differently than most regular Twitter users, biasing the analysis. The nature of our analysis also requires to focus on individual users, removing from our dataset Twitter accounts that are linked with company or professional profiles. These accounts are mainly used to advertise their owner and are clearly differentiated from Twitter accounts used by "regular" users [33,107]. Filtering individuals from a list of Twitter profiles is an open research problem that we aim to target in our future work. For the purpose of this work we randomly sampled 1,000 users, from our dataset, and manually marked the individual users. For this sample we evaluated a number of different profile features to identify the distinguishing factors for individual users. These features included the number of friends and followers, number and frequency of tweets etc. Our analysis showed that the cardinality of the intersection between the sets of followers and friends of a user is a satisfactory distinguishing factor for identifying individual users. Reciprocal relationships are also used to identify close friends [44], which is a characteristic of individual users. Based on this result we use this feature and remove all "corporate" and bot accounts from our dataset.

Collected data: Table 4.1 summarizes the collected data for each geographical area after data cleansing is performed. Overall we retrieved information for more

Name	Location	Users	Tweets	Geo-tagged Tweets
TW-NL	Netherlands	702,593	668,684,891	16,445,151
TW-LA	LA County	350,637	532,738,302	35,645,531
TW-LO	London	182,272	232,331,077	35,406,092

Table 4.1: Home location dataset: Number of users, number of Tweets and geo-tagged Tweets, for each of 3 regions of the resulted dataset.

Name	Post-code areas	Average area radius (*Km*)	Ground Truth Users
TW-NL	286	2,68	1414
TW-LA	62	2,75	370
TW-LO	151	2,37	760

Table 4.2: Home location dataset: Number of post-code areas and average area radius in Km, for each of 3 regions of the resulted dataset.

than 1 million Twitter users. This information contains around 1.5 billion Tweets, 6% of which contain geographical information. This number is significantly larger than most related work [48]. In all datasets we use Twitter API [1] following its terms of use with respect to users privacy.

Ground truth dataset: We used public information contained in Twitter user profile, manually inserted by the users, in order to create a ground truth dataset for evaluating our approach. Similar with [54] we assume that the location field in an OSN profile contains information regarding a user's home location. To this end, we search the profile information location field for exact geographical coordinates or user-reported post-code information. Then, we use either of these values to map the user to a post-code, considering that to be the user's Home location. Table 4.2 details the number of users contained in our ground truth dataset, for each area of interest. The table also lists the number of unique post-codes for which we have users and the average geographical area covered by each post-code. The latter value also constitutes the average granularity in which we can actually locate a user's Key locations.

Previous work dataset: To further strengthen the evaluation of our method

[1] `https://dev.twitter.com/rest/public` (Last accessed: June 2016)

and compare against state-of-the-art approaches we apply our methodology to the dataset retrieved by [31] and used by Yuan et al. [109]. This dataset includes geo-tagged micro-blogging activity and Home location ground truth for USA 9,475 users. We refer to this dataset as *GeoText* [2].

4.2.2 Workplace Location

In contrast to Home location, Work location is not usually clearly stated by a Twitter user in her personal profile. The reason for this is that Twitter profiles are used for a completely different purpose than career-related tools. LinkedIn on the other hand, is a professional social network where users publish career related information, including (among others), their current location and place of work.

To construct a work location dataset we use FriendFeed, an online OSN profile aggregator tool. FriendFeed allows its users to aggregate information posted into multiple OSNs by adding their profile accounts to a central service. For our dataset we collect FriendFeed accounts, whose owner have added both their Twitter and LinkedIn profiles, from FriendFeed's public stream during January 2015. We then used Twitter and LinkedIn APIs to retrieve the public profile information of the collected users, concluding to a list of 3,285 users. For these profiles we were able to collect both the geo-tagged activity of the user (Twitter) and the user's work location (LinkedIn). To the best of our knowledge, this is the first study that builds a dataset for user work location identification using OSN.

Data Cleansing: Despite the fact that the majority of LinkedIn profiles include information about a user's current employer, details regarding the exact geographical location of a company is limited. Additionally, when such geographical information is available, usually is related to the company's global headquarters and not the exact branch where users work at. For that reason we performed a pre-processing analysis in order to identify the exact branch of the company where a user works, along with its (self-stated) location at post-code level. As a first step we used users location field from her LinkedIn profile, that provides information about users' locations at city level. We then aimed to find the companies with the same name, as the one in the user's current employment field, in the area close to users reported

[2]Geo-tagged Microblog Corpus: `http://www.ark.cs.cmu.edu/GeoText/` (Last accessed: June 2016)

Name	Users	Tweets	Geo-tagged Tweets
TW-LinkedIn-Work	317	915,933	73,003

Table 4.3: Workplace location dataset: Number of users, number of Tweets and geo-tagged Tweets.

		Percentage
Country of origin	United States	34.7
	Great Britain	11.3
	Italy	5.7
	Spain	5.1
	Canada, France, Turkey	4.7 (each)
	Other Countries (23)	29.1
Industry	Internet	21.8
	Information Technology	16.4
	Marketing and Advertising	11.7
	Computer Software	8.2
	Online Media	7.6
	Other Industries (51)	34.3

Table 4.4: Workplace location dataset: Demographic characterisation

location. If the location is not identified we automatically discard the user profile from our analysis set. Users who do not include information about their employer were also discarded. Following this approach we managed to identify geo-location information for the workplace of 317 different users from different countries and map them to their corresponding post-code area. The resulted dataset has been manually inspected for validation.

Collected data: Tables 4.3 and 4.4 summarize the data collected for inferring users workplace locations. Our sample is multi-cultural as it contains users from a variety of countries of origin who are working in different industries.

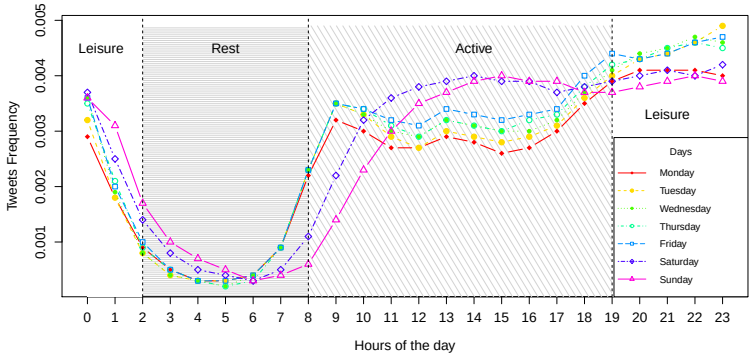


Figure 4.1: Tweets publishing activity during a week. Based on differences in behavior, day is divided in different time-frames. Rate is calculated divided by the total tweets quantity of the whole week.

4.3 Users Key Locations

Most previous work in user location identification from Twitter ignores two important observations that actually characterize users daily routine, not only in their online activity but also in their real life habits. These are: (*i*) users tend to spend a significant, but distinct, amount of their time during an average day in two key locations namely their Home and Work; (*ii*) these two locations are much more likely to appear in the user's geo-tagged activity during these specific timeframes, than locations that are not so frequent in user routine.

These observations are intuitive for users when considering our physical world interactions. Since we are interested in key location identification we use the ground truth dataset, described in [27], to evaluate whether these observations are also present in users' Twitter life. Figure 4.1 plots the percentage of Twitter activity (y-axis) for the different days of the week (lines) and the different time of each day (x-axis). We can clearly see the diurnal pattern in tweeting activity. Early morning hours show less activity than hours in the morning-afternoon and evening hours. Additional, we can observe the points in which user behavior seems to change, i.e. around 2 AM and 7-9 AM. [3] Furthermore, we can observe a slight shift in the tweeting activity of the users during weekends, as compared to weekdays. This shift denotes differences in the behavior of the user during weekends, an observation also made by Herder et al. [42], when analyzing user trajectories. Due to this observation we decide to ignore weekend activity when searching for the user's home and work

[3] Similar behavior has also been observed by [32]

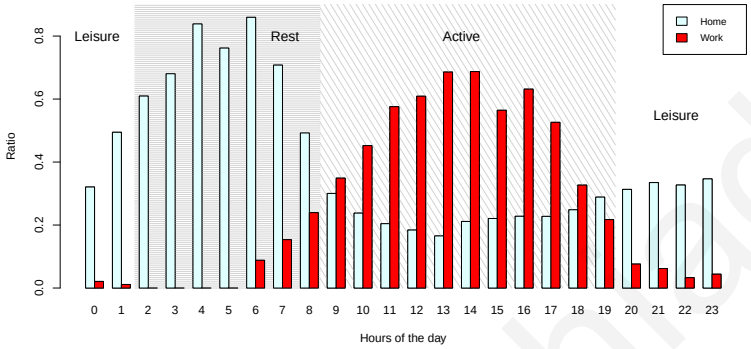
locations. We include this activity at a later state when we want to analyze the Leisure locations a user visits.

Based on the above observations, we argue that user's activity can be split in three different time frames related to the place that the user might be during that period. These timeframes are: (*i*) *Rest* time, between 2 and 8 AM, the time that the user most likely resides at her *Home* location (*ii*) *Active* time, during 8 AM and 7 PM, denoting the time that the user will most likely be at *Work* and (*iii*) *Leisure* time during the rest of the day, where the user spends her free time most probably outside the home and work environment.

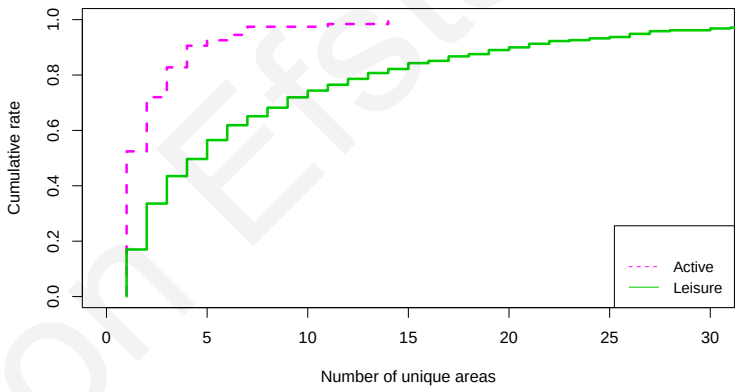
We expect that a user will mostly be posting tweets from a single location during the *Rest* and *Active* timeframes. Figure 4.2(a) examines this hypothesis for the *Home*(cyan) and *Work*(red) key locations. Using our ground truth dataset, for each case, we plot the ratio of user tweets sent from her reported home/work location during different hours of the day. The ratio is calculated as the fraction of tweets user u posted at each specific hour during the day from her home/work location over the total number of tweets of the user for that hour. As we can see from the results, the probability tends to increase significantly during (and close to) the *Rest* timeframe for the Home location, and during the *Active* timeframe for the Work location. Our observations also agree with the results of an analysis performed on a single user from Yuan et al. [109].

Figure 4.2(b) examines the number of different locations the user tweets from during *Active* and *Leisure* timeframes. We excluded the user's reported *Home* location from this analysis. We observe that in 90% of the cases the user will post, at max, from a handful of locations during *Active* timeframe. Having in mind that the user spends most of this time at her workplace, we expect it to be the most popular of these locations. The Figure also plot the CDF of different locations a user tweets from during the *Leisure* timeframe. The number of different locations is significantly higher in this case. Around 50% of the users tweet from more than 10 unique locations during this timeframe. This observation clearly demonstrates the different habits the users have in the different timeframes.

Discussion. Our analysis shows that the majority of users demonstrate temporal activity patterns on Twitter highly related with their home and work locations. By analyzing the geo-tagged information we can conclude that tweeting activity during *Rest* timeframe is more likely to be generated from *Home* location. During *Active*



(a)



(b)

Figure 4.2: (a) Ratio of tweets published from user's reported Home and Work locations on an hourly basis. Y-axis represents the portion of total geo-tagged Tweets that have been produced during the specific hour. (b) Number of different locations from which a user tweets during Active *and* Leisure *hours.*

timeframe activity is mostly likely to be generated from *Work* location. This result clearly indicates that actual information about a user's key locations can be inferred from Twitter activity.

4.3.1 Key Location Identification Model

The above observations verify our hypothesis that the user is much more likely to tweet from her *Home* location during *Rest* hours and from her *Work* location during *Active* hours. Based on these remarks, we define our key location identification method as follows: Given a set of geo-tagged tweets T_u of user u and the place P_u the tweets where posted from, we first split this set into three subsets, R_u, A_u and L_u containing the tweets during *Rest*,*Active* and *Leisure* timeframes respectively. We then estimate the *Home* and *Work* locations of the user by finding the most "popular" location during "non-working" (R_u and L_u) and "working" (A_u) hours, respectively. The popularity in each case is calculated as follows

$$W_u = \arg\max(\forall p \in P_u | t_p \in A_u : \sum_{i=day_1}^{day_n} A_u(i,p)) \tag{4.1}$$

$$H_u = \arg\max(\forall p \in P_u | t_p \in (R_u \cup L_u) : \sum_{i=day_1}^{day_n} w_r \times R_u(i,p) + w_l \times L_u(i,p) \tag{4.2}$$

Equation 4.1 calculates the most popular place, in number of unique days, among all the places the user tweeted during the *Active* timeframe, A_u. Equation 4.2 calculates the most popular place, also in number of unique days, among all the places the user tweeted during both the R_u and L_u timeframes. According to Figure 4.2(a) users tweet from Home with higher probability during R_u. To take this observation into account we apply a different weight w_r to the popularity of a place p if the tweet is included in R_u, and w_l if the tweet is included in L_u.

We calculate the weights by estimating the average, amongst all users, fraction of tweets from the home location over the total number of tweets during the two different timeframes. Table 4.5 shows the weight values for our three home location ground truth datasets. We can observe that the weights for all areas are almost identical. This shows that our method can easily be adapted to any area of interest

Dataset	*Rest*	*Leisure*
TW-NL	0.744	0.362
TW-LA	0.735	0.357
TW-LO	0.737	0.354

Table 4.5: Probability of tweeting from Home *during* Rest *and* Leisure *timeframes for the 3 different datasets.*

without changing the weights. We use the average of all three in the evaluation of our method.

4.4 Evaluation

We evaluate our key location identification method, proposed in the previous section, at post-code granularity both for Home and Workplace locations. For the Home location case, we evaluate our method using two different approaches. First, we compare the identified user Home locations with the user reported home location, as extracted from the user's profile entry. Second, we proceed to a comparison of our results with publicly available socio-economic data. In specific, we compare the post-code population density in Home locations, with the ones that we derive by applying our method in our Twitter dataset. Furthermore, we compare the estimated workplace locations against the exact workplace locations identified both from LinkedIn and Twitter data.

Metrics and Methods

We validate our approach based on well established metrics used in literature. These are:

ACC *Accuracy* gives the percentage of correctly inferred users' key locations over the total sample size [50,67,91].

ACC@R *Accuracy within radius* (R) gives the percentage of correctly inferred users' key locations identified within R Km from users reported locations [50,67,91].

AED *Average Error Distance* defines the distance, in Km, between the inferred location (center of the post-code in our case) and user's reported location [50,91].

Using the above metrics we evaluate our method and compare it with the state-of-the-art geo-tagged data user location methods as those are defined in related work. These are:

MP *Most Popular* marks as home location the most popular location, in number of geo-tagged tweets, visited by the user [35].

MC *Median Clustering* marks the user's home location by calculating the median value of location the user tweeted from [91].

TF-C *TimeFrame - Clustering* is the method proposed in this work. The method takes into account the fact that the user usually resides in different locations during different times of the day and week.

4.4.1 Home Location identification

Data pre-processing

Before applying our method to either dataset we first do a pre-processing pass over the data, to eliminate common well known locations and bring all geo-tagged information to a common format at post-code granularity. Popular locations are referred in Twitter as *Points Of Interest* (POI). These locations define specific attractions, local businesses, landmarks etc. POIs are not used to define a user's home place, and for this reason we decide to remove such places, marked with a specific tag in the tweet location field, from the user's Twitter stream.

In a second step we map geographical coordinates contained in the tweet location field to the closest post-code in terms of euclidean distance. We choose postcode level over other forms of mapping, i.e. city or arbitrary geographical boundaries [4], since it is a well defined and official boundary on one hand and much more precise on the other.

Evaluation with ground-truth data

Results. Table 4.6 presents the evaluation of TF-C in correctly identifying the *Home* location of the user, for the three different geographical locations, along with the comparison with the aforementioned state-of-the-art methods. Overall TF-C outperforms the other methods, in both metrics presented in the table. In terms of

[4] Cho et al. [17] used a 25Km square boundary

Method	TW-NL	TW-LO	TW-LA
	ACC		
MP	0.69	0.47	0.55
MC	0.67	0.19	0.39
TF-C	**0.81**	**0.68**	**0.701**
	AED		
MP	3.21	4.13	6.05
MC	3.93	5.21	8.15
TF-C	**2.77**	**2.05**	**2.63**

Table 4.6: Home-Location identification performance Accuracy (ACC) and Average Error Distance (AED) in Km, of the compared approaches in 3 different areas.

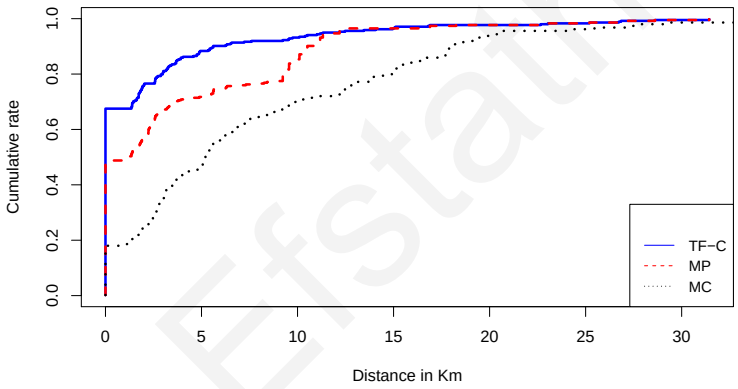


Figure 4.3: TF-C performance for London dataset. Proposed methodology is able to identify the exact post-code location with 68% accuracy and performs better in lower granularities than compared approaches.

accuracy TF-C can identify more than 80% of the user's home locations, in the country of Netherlands, while in any case it can identify more than 70% of the user's home. In comparison with the other methods, TF-C performs 20-50% more accurate.

In terms of the AED metric we can see, from Table 4.6, that TF-C locates the user closest to her Home location, with values always being less than 2.7Km from the center of the user defined post-code. Recall, from Table 4.2, that the average area radius for the post-codes in our dataset is also around our method's AED values. All other methods identify the user at least 3.2Km from her defined location, and in some cases reach error distances close to 8Km.

Figure 4.3 compares the evaluated approaches in terms of the ACC@R metric

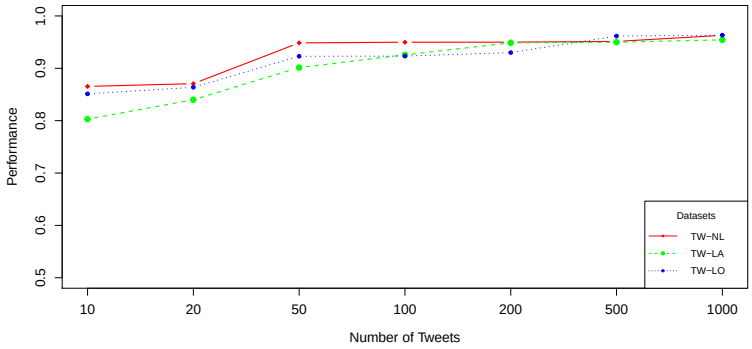
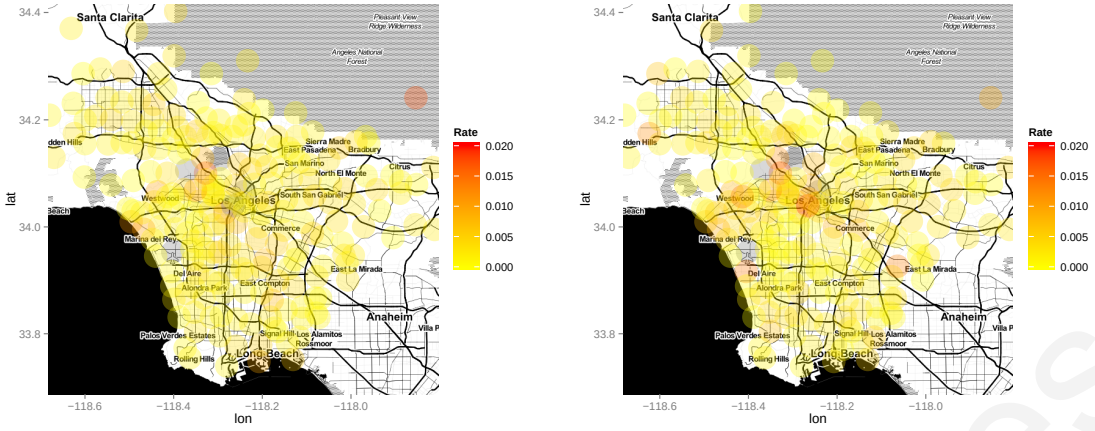


Figure 4.4: Performance of proposed method in contrast to the number of recent tweets for the 3 datasets.

for the TW-LO dataset. The figure plots the total accuracy of each method as a function of the distance from the center of the user defined postcode. From the results we observe that *TF-C* can identify more than 95% of the users in less than 10 Km from their center location, and more than 80% in less than 5Km. The *MP* and *MC* methods reach the same level of accuracy (80%) for radius larger than 10 and 15 Km respectively. Also, *TF-C* can identify all users in less than 20Km, versus the 30+ Km of the two comparison methods.

The proposed methodology is able to identify users key locations using geo-tagged tweets. However, as presented on Table 4.1, only 6% of the total collected tweets contain meta-data regarding their location. Figure 4.4 examines the number of tweets needed, by our method, to accurately identify the user's Home location. As we can see from the figure, 10 to 20 tweets are enough for TF-C to identify more than 85% of all identified users. Recall, that TF-C is able to accurately infer a user's key locations when her tweeting activity follows a distribution similar to the ones presented in figure 4.2(a). The work of Sadilek et al. [92] required at least two of the user friends to have at least 100 geo-tagged tweets, a number much larger than our approach. [51] provides a comparison of accuracy of a number of different location prediction models (Table 4 in that work); Most Popular (MP) being one of the models examined. Their work shows that all state-of-the-art algorithms require at least 100 tweets from each user to provide a prediction accuracy of 72%, at 30Km granularity in the best case senario. With the above numbers in mind it is clear that our approach can provide higher accuracy, at post-code granularity, for both new and old twitter users, using only a small amount of their tweet activity.



(a) Differences between real and predicted population rate.

(b) Differences between real and predicted employees rate.

Figure 4.5: Predicted population was calculated after applying the proposed model on a dataset of 350,000 users from LA county. Real population was collected from LA county's official statistics.

Discussion. Results show that TF-C outperforms the state-of-the-art in geo-tagged data based key location identification methods by at least 15% and up to 50% in terms of accuracy. Also our method can detect user's home location in a radius smaller than 10Km in most of the cases. *MP* and *MC* are both methods used to provide ground truth data for social community based [92] and content-based [67, 91]. All these methods result in low detection accuracy, between 20 and 70%, and also detect users in a much higher radius, more than 10Km in all cases. Our results show that TF-C provides a more accurate ground truth for user's home location, that will help improve both the methods themselves and their detection accuracy. In future work we plan to both evaluate our approach against such methods and quantify the improvements a better ground truth dataset can provide.

Evaluation over previous work dataset

We also evaluate our approach over the *GeoText* dataset, collected and used in previous work related to user location identification. Home location of each user in this dataset is already provided by Eisenstein et al. [31]. Our evaluation results show TF-C identifies the home location of the users in this dataset with an accuracy of 76%. Yuan et al. [109] also used the above dataset for evaluating a user location identification method based on the tweet text. Their approach uses training and prediction of the user location and gives prediction accuracy significantly lower than TF-C.

Comparison with open-data

Results. In the previous section we evaluated the accuracy of our method and demonstrated the improvement it offers over the related work. In this section we use open data from the County of Los Angeles to derive the population of each different post-code as a function of the total population of the County. Figure 4.5(a) shows a heat-map of the differences in the population distribution derived from the real data compared with the population distribution as this can be derived by our method, for 200,000 Twitter users. As depicted in the heatmap, for about 87% of the areas the predicted and real post-code population rate differ only by 0.005.

Discussion. Nowadays, the population census procedure is performed with the use of well studied and applied methodologies, like door-to-door interviews at a sample of habitants. Despite the fact that these enumeration methodologies provide us with accurate data, they do have several limitations. [5] Such limitations are the cost of performing such a study, the time needed for its completion and the access to the sample that will be used. Thus, such demographic studies take place on a 'several years' base and usually are out-dated. Based on the accuracy provided by our methods, we believe that *TF-C* can act as a complementary and closer to real-time method for performing demographic studies. Using data available from OSNs one can quickly and in zero cost get a close to real estimate of the current trends in an area of interest, without waiting for the more complicated population census procedure.

4.4.2 Identifying workplace location

In this section we proceed and evaluate our approach' accuracy in predicting a user's *workplace location* based on her interactions in Twitter. To the best of our knowledge, this is the first study where geo-location information about workplaces has been collected and used for such an analysis.

Data pre-processing

We use the LinkedIn-Twitter dataset described in section 4.2.2 for this evaluation. Contrary to the home location evaluation case, we do not remove popular locations, referred by Twitter as Points Of Interest (POI), from the workplace evaluation. These attractions or local businesses were removed from the previous analysis as they are

[5] `https://www.census.gov/prod/1/gen/95statab/app3.pdf` (Last accessed: June 2016)

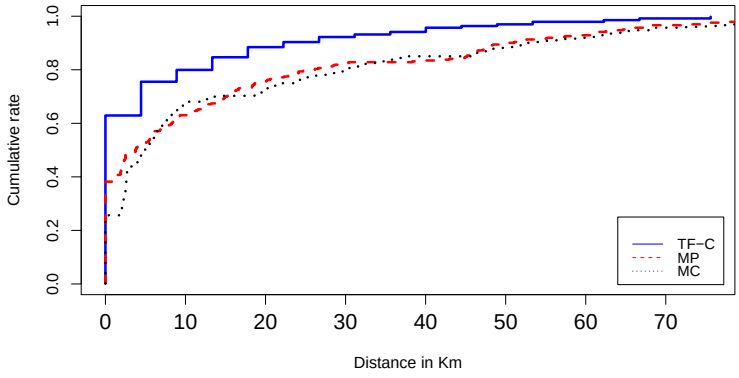


Figure 4.6: TF-C performance for identifying workplace location from a global dataset. Proposed methodology is able to identify the exact workplace location at post-code granularity with 63% accuracy.

not used to define a user's home. However they could represent a user's workplace.

Similarly with home location identification, we map geographical coordinates contained in tweet location field to the closest post-code area. However, because we use a world-wide dataset and we do not have access to global post-code information, we divide the global geographical space in boundaries with radius equal to 2Km, which is less than the average post-code coverage size in Netherlands, London and LA county. We then map each tweet to the corresponding boundary area.

Evaluation with ground-truth data

Results. Figure 4.6 compares the evaluated approaches in terms of the ACC@R metric for the global workplace dataset. The figure plots the total accuracy of each method as a function of the distance from the center of the user workplace postcode. As we can see from the results, our method is able to detect a user's workplace location with similar performance as her home location. Specifically, it is able to detect the exact post-code location with an accuracy of 63%, in comparison with *MP* and *MC* methods which have 38% and 26% respectively. Additionally, in a 10Km radius, our method, is able to identify the employers location for more than 80% of the total sample. *MP* and *MC* methods both reach the same level of accuracy in a much larger radius of about 40Km. Also, *TF-C* method is able to identify more than 90% of the users workplace location in a radius smaller than 20Km, while both *MP* and *MC* need a radius of more than 50Km to reach similar level of accuracy.

Discussion. Our results demonstrate that *TF-C* achieves high accuracy in workplace location identification, on a worldwide dataset, at a granularity equal to a post-code area. From these results we can see that information about a user's workplace area can be derived from public data, despite the fact that she does not explicitly reports it. In this work we take into account only the meta-data of users activity in Twitter, taking advantage of the fact that our interactions in Online Social Networking platforms sometimes generate more information than the one we intend to share.

Comparison with open-data

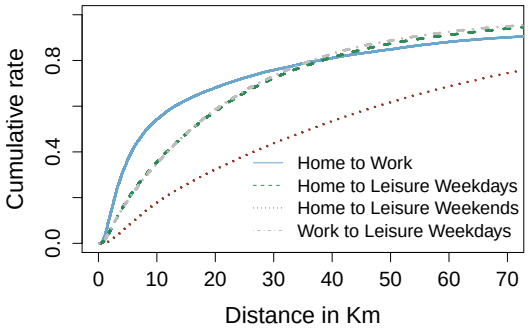
Results. After identifying the workplace location at post-code granularity of a sample of users in Los Angeles county, we proceeded in comparing the general statistics with open-data collected from this area. Figure 4.5(b) presents the differences in the rates between real and predicted employees fraction over total employees of each post-code area. As we can see more than 85%, of post-code areas differ by less than 0.005, while 5% differs by more than 0.01.

Discussion. Our results show that TF-C is able to provide insights to real-world studies that are more complex than population census. Methodologies that are being applied in such studies are well validated and commonly accepted, however, the identification of users key locations from their online social networking activity can also help in this effort.

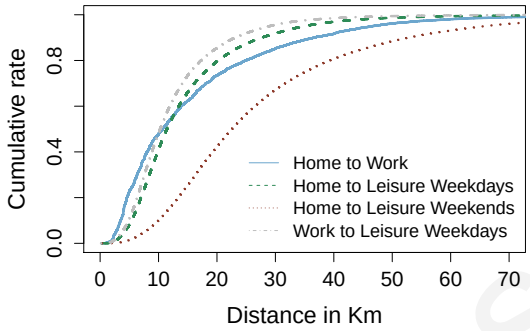
4.5 The Location Factor

Our methodology, presented in the previous sections, provides a more accurate method for identifying Twitter users Key locations. In this section we scratch the surface of the location factor in both the user's daily mobility patterns and the formulation of a user's social network. We target to answer the following questions:

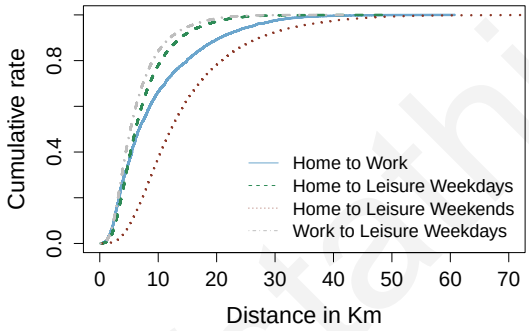
- How are the daily mobility patterns of the users affected by their Home and Work locations? (Subsection 4.5.1)
- How is the user Ego network formed based on the location of the user? (Subsection 4.5.2)
- How the user's Home, Work and Leisure locations affect her sentiment? (Subsection 4.5.3)



(a) The Netherlands



(b) Los Angeles county



(c) London

Figure 4.7: Cumulative Distribution Function (CDF) of the distances between Home, Work and Leisure locations for the three geographical areas of our dataset. The distance is measured as the absolute distance in Kilometers from the Key location of reference.

4.5.1 Daily mobility patterns

Home-Work-Leisure proximity: Figure 4.7 plots the CDF of distances between Home, Work and Leisure locations for the three geographical areas of our dataset. The distance is measured as the absolute distance in Kilometers from the Key location of reference. The figure shows an obvious tendency of the Dutch (Figure 4.7(a)) and LA (Figure 4.7(b)) Twitter users to choose Homes further away from their Work location compared to the Twitter users of London (Figure 4.7(c)). Only about 50% of the Dutch and LA Twitter users live in less than 10Km from their Work location. In the case of London Twitter users the Home-Work distance is less than 10Km for more than 60% of the users. The London curve also grows faster showing more than 85% of the user's to reside in less than 20Km from their Work location. In the

case of The Netherlands the tendency to live further from Work can be explained by the standard allowances, in the form of tax deductibles, for workers commuting in distances more than 10Km between Home and Work [6].

Figure 4.7 also plots the distances of the selected Leisure locations where Twitter users of the three areas spend their Leisure time, relative to their Home and Work locations. We can see that both LA and London Twitter user's spend their Leisure time closer to their Home and Work locations during weekdays. 90% of London users travel less than 10Km for an after work drink. Similarly, LA Twitter users mostly travel up to 20Km during a workday. Users from The Netherlands show a different behavior, only 50% travels less than 20Km during a week day for Leisure activities, with the rest 50% of the cases traveling distances even longer than 70Km. During weekends the traveling habits change in all three case studies. Twitter users travel a much longer distance for Leisure activities. London users still travel the shortest distance from Home during weekends, while users from The Netherlands seem to make significantly longer travels. Furthermore, the average distance traveled from Home to Work locations for the city of London is 13.4Km, which shows high accuracy compared to 14.6Km reported in open-data [7].

Figure 4.8 further examines the distance traveled by the Twitter users of our three case studies for Leisure purposes. The figure plots the fraction of Leisure locations as a function of the distance traveled from the user's Home location. Cho et al. present a similar study using *cell phone location data*, *Growalla* and *Brightkite* check-ins [17]. Likewise, we also observe a change in the slope of the curve around 100Km. From both results we can conclude that geo-tagged Tweets distance from Home follows similar distribution as the three studied cases. All three locations show an almost identical distribution up to the point of 100Km. The distribution decays faster up to that point and flattens for distances longer than 100Km. This effect is more obvious in the case of The Netherlands, where we see an almost flat line for distances longer than 100Km and shorter than 1000Km.

[6]Organization for Economic Co-operation and Development, `http://www.oecd.org/social/soc/47346594.pdf` (Last accessed: June 2016)

[7]London DataStore, `http://data.london.gov.uk/` (Last accessed: June 2016)

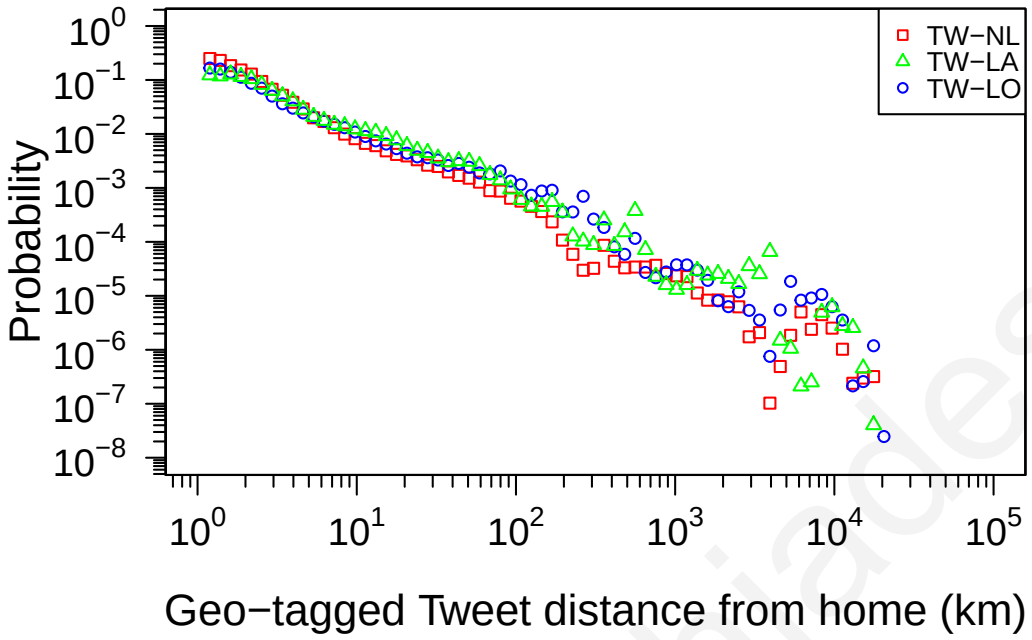
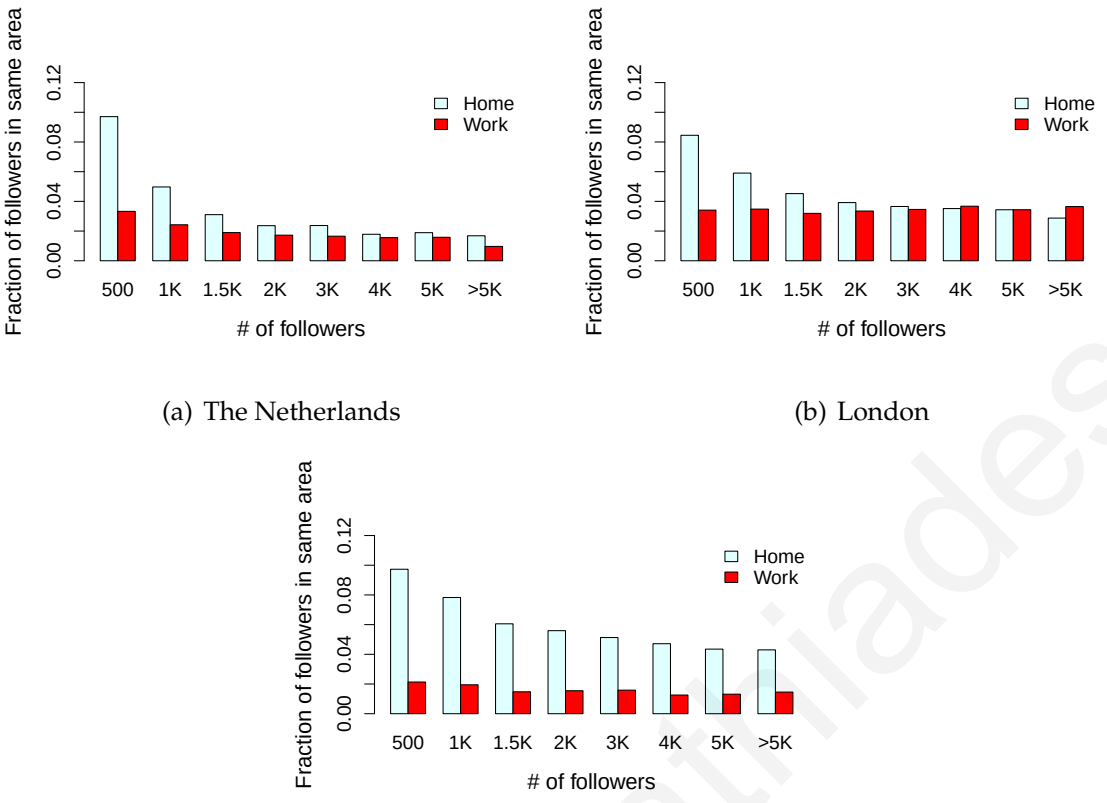


Figure 4.8: Fraction of Leisure locations as a function of distance traveled from Home location.

4.5.2 Social Network formulation

Next we examine how a Twitter user's Key locations function into the formulation of the user's Ego network. We first look at the distance of the user's followers from the user's Home location. Secondly, we examine the distance of strong relationships as these can be identified from the reciprocity factor. Finally, we use the open data available for the LA county to examine the demographic relationship of the user's connections.

Local Vs. Global connections: Figure 4.9 plots the percentage of the user's followers that live/work at the same location as the user as a function of the user's total number of followers. Home (cyan) bars shows the percentage of user's followers that live at the same location (same postcode) as the user's Home location. Work (red) bars show the percentage of user's followers that work at the same location as the user's Work location. In the case of Home-proximity followers we observe a descending trend as the number of followers of the user increases. This trend is obvious in all three geographical areas. In the case of Work-proximity the percentage of followers in the same location seems not to be affected by the number of followers of the user. We observe similar percentage of followers in the same location despite



(a) The Netherlands

(b) London

(c) LA county

Figure 4.9: Fraction of user followers in the same Home location as the user as a function of the user's total number of followers.

the number of followers of the user. Overall, we can observe that user's mostly have a larger number of followers from the same Home location than the same Work location.

Figure 4.10 plots the fraction of a user's followers as a function of their distance from the user's Home location. We can observe that the majority of a user's followers live in close proximity to the user. This effect is more obvious in London (blue bars) and LA county (green bars) where more than 90% of the user's followers live in less than 50Km from his Home location. The Netherlands (red bars) show a slightly different approach where only 50% of the users followers are in a less than 50Km proximity from the user's Home location. Additionally, we observe an increase in the number of followers located more than 100Km from the user's Home location.

Strong connections spatial distance: Reciprocity examines whether a connection in a directed graph is bidirectional. That is, if a user A follows a user B and user B also follows user A we define their relationship as reciprocal. Reciprocity is often

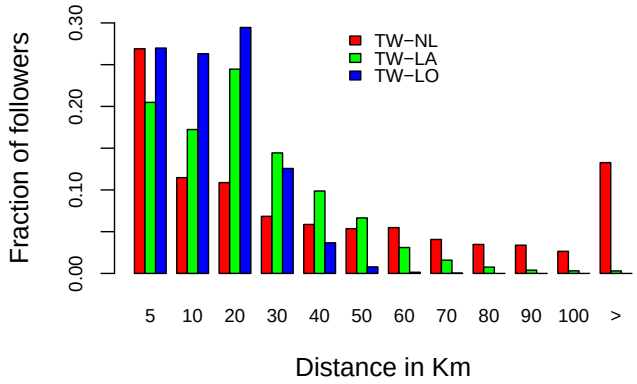


Figure 4.10: Fraction of Twitter user followers as a function of the distance from the user's Home location.

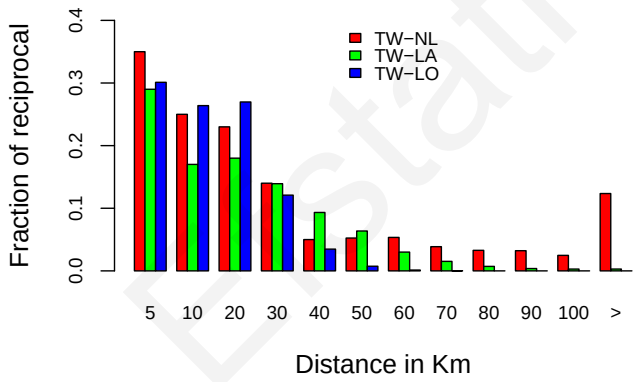


Figure 4.11: Fraction of reciprocal relationships of a Twitter user as a function of the distance from the user's Home location.

consider as a measure of a stronger connection between two users [2]. Figure 4.11 plots the fraction of reciprocal relationships of a Twitter user as a function of the follower's Home distance from the user's Home location, for all users in the three datasets. Again, we observe similar results as in Figure 4.10. Most of the reciprocal connections of a user are close to her Home location, in distances less than 50Km. Reciprocal relationships usually show an underlying relationship between the two users, outside of the online social network. Our results also strengthens this believe since most reciprocal relationship are in a proximity that allow face-to-face interaction.

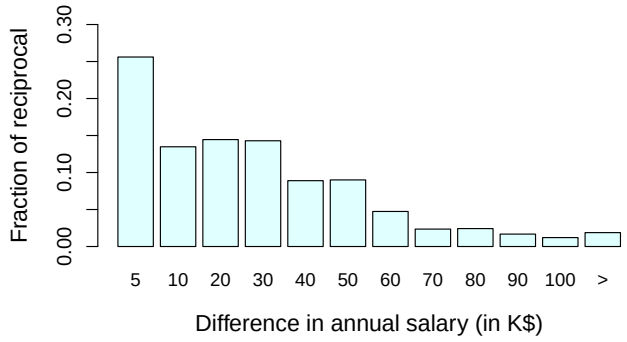


Figure 4.12: Fraction of reciprocal relationships of a Twitter user as a function of the difference between the average annual salary of the two postcode areas.

Nodes proximity: To further examine these reciprocal relationships we compare the similarity of these connections as it can be defined by the use of open data demographics. We use the available open data for Los Angeles county and examine whether these reciprocal connections belong to the same economic status of the user. We define the economic status as the average income salary for the postcode. Figure 4.12 plots the average fraction of reciprocal connection as a function of the difference with the user's postcode average income. We can see that most of these connections tend to live in a postcode with similar average income. With this result in mind we can say that users not only tend to be friends, in Twitter, with users residing in close proximity but also tend to be friends with other users of the same economical status. These results could be used for improving the approaches that address the *link prediction problem*, as we conclude that the economical status influences the "proximity" of the nodes in Twitter network [64].

4.5.3 Sentiment

Sentiment is commonly used to measure the emotions of user's natural language. It can show the reaction of users to several events or their emotional state during a conversation. Combined with location information it can show how different geographical areas react to specific events or express themselves during their everyday online interactions. For example, Hedonometer is a tool used to measure the average happiness of Twitter users, also segmenting the tweets to the different US states they originate from [8]. Our method, able to identify the actual Home and Work locations

[8]Hedonometer, `http://hedonometer.org/index.html` (Last accessed: June 2015)

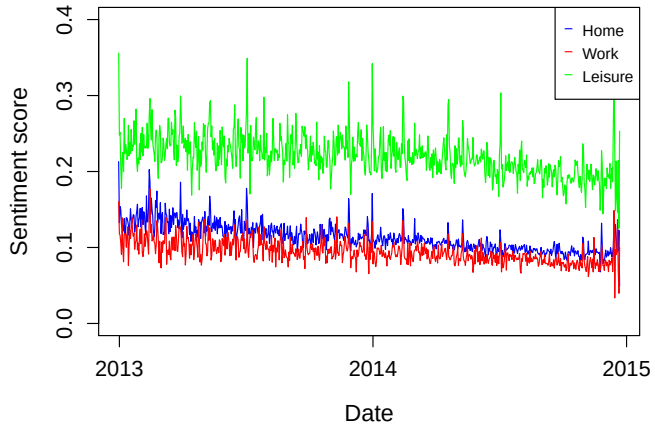


Figure 4.13: Sentiment per calendar day for the Tweets published from Home, Work and Leisure areas.

of the users, can be used to zoom in into the different neighborhoods and examine the sentiment of the different Key locations of the users.

To illustrate such an example Figure 4.13 plots the sentiment score for each calendar day from January 1st, 2013 to January 1st, 2015 for the three different categories of *Key locations*. We use the Python Natural Language processing ToolKit (NLTK) [8] to identify sentiment for each English tweet in our dataset, due to the fact that it has been widely used for Tweets sentiment analysis [80,108]. NLTK analyzes the Tweets sentiment using text classifiers trained on twitter and reviews datasets [9]. We segment the tweets in three categories, Home, Work and Leisure, based on the *Key locations* identified for each user. For each day, we calculate the average score of each user for the corresponding category. Then for each category we calculate and plot the average score of the specific calendar day. As we can see, users tend to publish more positive Tweets from Home than from their workplace. The most positive locations during the whole period of investigation were published from Leisure locations. In specific, Tweets that are published from Leisure locations constantly have sentiment scores two times higher than the ones published from Home locations. Furthermore, we do not observe any change in the trends of sentiment. "Happy" and "Sad" days can be identified in all timelines, with tweets from Leisure locations being constantly more positive.

[9] `http://text-processing.com/demo/sentiment/` (Last accessed: June 2016)

Chapter 5

Sentiment of Entrepreneurs in Twitter

Our results highlight the influence of users key location on their sentiment in the content that they post. Based on our findings, we turn our attention on examining other factors that could potentially influence the sentiment of a user. For this study we focus on a special category of people, *entrepreneurs*, and we examine their sentiment on the content that they post on Twitter. We compare our findings with the general population, but also across the different entrepreneurship types.

5.1 Hypothesis

We organized our work into four hypotheses:

H1: *Entrepreneurs are more likely than non-entrepreneurs to exhibit positive general sentiment.*

An entrepreneur is well-informed about their own preferences, and motivated to achieve them. As a result, they are more likely to feel greater motivation and well-being while undertaking their duties [90]. Psychology suggests that when a person is involved in activities he consider valuable, he has positive emotions from the engagement [14]. We believe that the positiveness is reflected also in their daily routines and their OSN interactions through the content that they share with their followers.

H2: *Entrepreneurs are less likely than non-entrepreneurs to exhibit positive sentiment in respect of business matters.*

Despite the fact that entrepreneurs should be overall more positive than the average user, we believe that this does not hold when the discussion comes to business related

topics. An entrepreneur is able to identify specific threats and high possibilities for failures due to his solid background and understanding.

The fact that entrepreneurship presents many difficulties and challenges can make an entrepreneur's sentiment less positive in relation to business matters. Furthermore, entrepreneurship has a high rate of failure, especially during the first few years [3, 104]. Faced with a real, personal threat, an entrepreneur's sentiment will typically be less positive when dealing with business matters rather than other matters where they are less personally exposed to risk. Thus, entrepreneurs are likely to feel less positive sentiments towards business than towards other matters.

H3: *Social entrepreneurs are more likely than other entrepreneurs to exhibit positive general sentiment.*

We next examine the sentiments of social entrepreneurs. Social entrepreneurs apply business practices to address social problems or create social value [22]. We believe that there are several reasons why social entrepreneurs' sentiment may be even more positive than traditional entrepreneurs. A social entrepreneur usually has personal participation in a social enterprise. They find their personal involvement in the activities more important than the outcome. This "warm-glow" pleasure is associated with observable changes in parts of the brain associated with reward [39]. Thus, we expect to see social entrepreneurs' warm-glow pleasure reflected in their sentiments.

H4: *Serial entrepreneurs are less likely than other entrepreneurs to exhibit a positive general sentiment.*

We argue that entrepreneurs who establish multiple businesses in sequence over time [86], serial entrepreneurs, may be less likely to show positive sentiment than other entrepreneurs. We start by proposing that a serial entrepreneur's experience leads to memories of entrepreneurial failures that lower their sentiment directed towards business matters. Thus, we believe that a serial entrepreneur who experiences less pleasant emotions than less experienced entrepreneurs are likely to show sentiments that are less positive.

5.2 Dataset

Our dataset is derived from messages and personal profiles placed on Twitter by entrepreneurs and non-entrepreneurs from January 2013 to January 2015. Twitter has

Dataset	*Initial tweets*	*Users*	*Entrepreneurs*	*Final usable tweets*
Los Angeles	631,738,302	350,637	4,062	4,590,538
London	232,331,077	182,272	1,500	2,152,140
Worldwide	12.1B	34.6M	25,180	22,833,489

Table 5.1: Number of users, initial tweets, and usable tweets.

become an important communications channel for entrepreneurs, investors, managers, and professionals so as to improve personal branding as well as to gather information, collect surveys and feedback, manage online reputation, track the information of competitors, and so on. Moreover, entrepreneurs use Twitter to contact or follow venture capitalists.

We created a database constructed of tweets and users public profile information. Our sample data set takes into account only tweets of individual users, both entrepreneurs and non-entrepreneurs, from their personal accounts. This means that we have to remove from the dataset tweets sent by Twitter accounts that do not correspond to individual users, and which could bias our analysis. These accounts are either bots (software that autonomously performs actions such as tweeting, retweeting, liking, following, unfollowing, or direct messaging other accounts) or are linked with company or professional profiles, which are mainly used to advertise their owner and are clearly differentiated from the accounts of individuals [33, 107]. To remove these accounts, we evaluated a number of different profile features (including the number of Twitter friends and followers, the number and frequency of tweets, and reciprocal relationships) which have been studied in the literature and act as the key factors for distinguishing individual Twitter users from other users [44]. We then used these characteristics as the basis for exclusion of non-individual users.

Overall, we retrieved information for about 36 million Twitter individual users and around 13 billion tweets. For our worldwide data, we identified 25,180 entrepreneurs, while for our London data we identified 1,500 entrepreneurs and for our Los Angeles data we identified 4,062 entrepreneurs (see Table 1). We also randomly selected three samples of individual non-entrepreneurs from the geographical regions with the same sample sizes, to be used for analysis and comparison in the remainder of the chapter. This random selection ensures that the analysis is numerically tractable.

Attribute (mean,stddev)	Sentiment	Entre-preneur	Followers	Following	Source	Follower fol-lowee	#tweets	Retweet	Geo-tagged
Sentiment (0.33,0.08)									
Entrepreneur (0.5,0.5)	0.3557****								
Followers (1545,2529)	-0.005	0.0067							
Following (639,2670)	0.0685*****	0.0136*	0.0688*****						
Source (0.45,0.49)	0.0696*****	-0.005*	0.0054	0.0126*					
Follower followee (5.61,440.65)	0.0106*	0.0104*	0.4878*****	-0.0114*	0.0215*				
#tweets (6051,12773)	-0.0217*	-0.0079	-0.0096*	-0.0072	-0.0225*	0.0018			
Retweet (0.22,0.41)	0.0134*	0.0061	-0.0111*	-0.0112*	0.0294**	-0.0103*	0.0413***		
Geotagged (0.026,0.019)	-0.0124*	-0.0067	-0.005	0.0189*	-0.0098*	0.0142*	-0.0058	0.009	
Hashtag count (0.32,0.81)	0.0195*	0.0163*	-0.0248*	0.0074	0.0059	-0.0204*	-0.0053	-0.0065	0.0013

*Table 5.2: Descriptive Statistics and Correlations for London Dataset. ($p < .0001$ '*****', $p < .001$ '****', $p < .01$ '***', $p < .05$ '**', $p < .1$ '*')*

5.2.1 Variables

Sentiment: Sentiment was measured using the VADER (Valence Aware Dictionary for Sentiment Reasoning) classifier library of NLTK for the sentiment analysis of the tweets. VADER is a lexicon and rule-based sentiment analysis tool that is tailored to specifically detect sentiment expressed in social media. VADER takes as input the posted text, emoticons and hashtags of tweets for building the training set and returns a score in the range of [-1, 1]. This score indicates the positivity or negativity of a tweet. Any tweet has an output score bigger than 0 is defined as a positive tweet, whereas any tweet that has an output score of less than 0 is defined as a negative tweet; any tweet has an output score equal to 0 is defined as a neutral tweet.

Tweet topic: In order to extract the discussion topic for each tweet, we submit every tweet separately to AlchemyLanguage API via an online HTTP REST Web

Attribute (mean,stddev)	Sentiment	Entre-preneur	Followers	Following	Source	Follower fol-lowee	#tweets	Retweet	Geo-tagged
Sentiment (0.3017,0.13)									
Entrepreneur (0.5,0.5)	0.2582*****								
Followers (4209,14167)	0.0048	0.0022							
Following (1752,1304)	0.0610*****	0.0895*****	0.0306***						
Source (0.53,0.47)	0.0592*****	-0.011*****	0.0095*	0.0360****					
Follower followee (3.65,203.17)	-0.0016	-0.0166*	0.8839*****	-0.004	0.0058				
#tweets (6438,15473)	-0.0165*	-0.0041	-0.0159*	-0.0104*	-0.0161*	-0.0085*			
Retweet (0.24,0.43)	-0.0169*	-0.0289***	0.0171*	0.0109*	-0.0007	0.0161*	-0.0145*		
Geotagged (0.03,0.011)	-0.0190*	-0.0208*	-0.0115*	0.0099*	-0.006	-0.0066	0.0089*	0.0049	
Hashtag count (0.22,0.67)	-0.0181*	-0.0019	-0.0113*	0.0143*	-0.0080*	-0.0117*	-0.0044	0.0113*	0.0064

*Table 5.3: Descriptive Statistics and Correlations for Los Angeles Dataset. ($p < .0001$ '*****', $p < .001$ '****', $p < .01$ '***', $p < .05$ '**', $p < .1$ '*')*

service. Alchemy API is a core component of IBM's Watson Developer Cloud . For any given text, AlchemyLanguage API returns the topic with the highest probability that the specific tweet belongs to. Technically, AlchemyLanguage categorizes unstructured text into a hierarchical taxonomy using custom annotation models. In our study, we ended up with 11 categories (arts and entertainment, business, computer and internet, culture and politics, gaming, health, law and crime, recreation, religion, science and technology, sports). In total we have classified 302,862 and 261,018 tweets posted by entrepreneurs and non-entrepreneurs respectively during the period September - October 2014.

Entrepreneur: Entrepreneurs are identified as users who have in their personal Twitter description any of the following terms: "entrepreneur", "founder", "co-founder", "business-owner", "business owner", "start-up", or "start up". Social and serial entrepreneurs are identified as users who describe themselves as such in their

Attribute (mean,stddev)	Sentiment	Entre-preneur	Followers	Following	Source	Follower fol-lowee	#tweets	Retweet	Geo-tagged
Sentiment (0.2987,0.083)									
Entrepreneur (0.5,0.5)	0.278*****								
Followers (493,4577)	0.069**	0.096*****							
Following (167,179)	0.08*	0.106*****	0.315*****						
Source (0.24,0.42)	0.0609*****	0.021*****	0.057*	0.049					
Follower followee (7.38,320)	0.005*	0.016*	0.816*****	-0.014*	0.009*				
#tweets (1449,8305)	-0.001	-0.039*	0.001	0.045*	0.002*	0			
Retweet (0.18,0.38)	-0.02*	-0.035*	0.021	-0.005*	-0.002	0.038*	0.02*		
Geotagged (0.028,0.016)	-0.006*	-0.016*	-0.026	-0.027*	-0.019*	-0.033*	-0.014	-0.016	
Hashtag count (0.33,0.81)	-0.014*	0.006*	0.036*	0.039*	0.038	0.021*	0.042	0.024	-0.036

*Table 5.4: Descriptive Statistics and Correlations for World Wide Dataset. ($p < .0001$ '*****', $p < .001$ '****', $p < .01$ '***', $p < .05$ '**', $p < .1$ '*')*

profile. As we identify entrepreneurs based on the English language versions of these terms, it is possible that many entrepreneurs outside of the English-speaking world who describe themselves as entrepreneurs in other languages will not be recorded as entrepreneurs in our database. To allow for this possibility influencing our results, we also consider tweets sent from two specific geographic regions where English is the dominant language, namely London in the U.K. and Los Angeles in the U.S. These two regions are moreover worthy of special examination as they are among the most successful in the world at attracting startups, corporates, and venture investors , and are leading centres of innovation in Europe and the United States.

Followers: Followers indicates the number of other users who follow the user. In Twitter, someone can choose to follow a user, meaning that they receive all of the user's messages. Thus, followers measures the popularity of the person's messages.

Following: Following indicates the number of the number of other users that the user follows. It is a measure of a person's interest and integration in Twitter.

Retweet: Retweet is a dummy variable that indicates if the specific tweet is a retweet. This means that the users is resending a message that was originally sent on Twitter by another user.

Geotagged: Geotagged is a dummy variable that indicates if the specific tweet is geotagged or not. If the tweet is geotagged then it has the specific latitude and longitude of the location from which it was sent.

Hashtag count: Hashtag count is the number of hashtags the tweet contains. A hashtag is a short word or phrase which other users can search for, in order to easily find messages carrying the hashtag. Hashtag count measures how easily people can find and read the tweet.

Source: Source indicates the application that was used for the generation and publication of the tweet. For example, if a user published the tweet using his Android smartphone, this attribute will have the value of 1 and 0 otherwise.

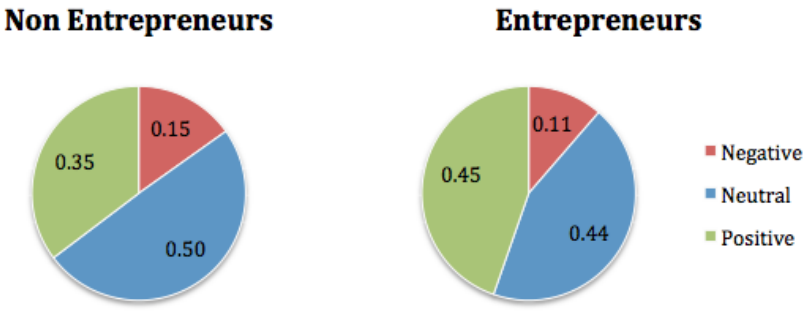
Descriptive statistics and correlations are reported on Tables 5.2, 5.3, 5.4.

5.3 Results

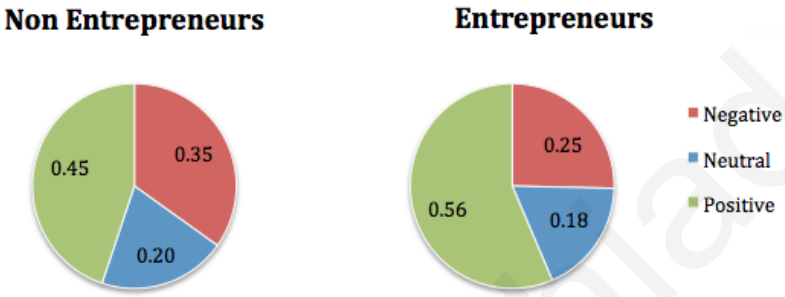
In this section we present the results of our analysis. The section examines the four hypotheses stated in Section 5.1 and provides evidence for their verification or not based on the three different datasets we examined.

H1: Entrepreneurs are more likely than non-entrepreneurs to exhibit a positive general sentiment

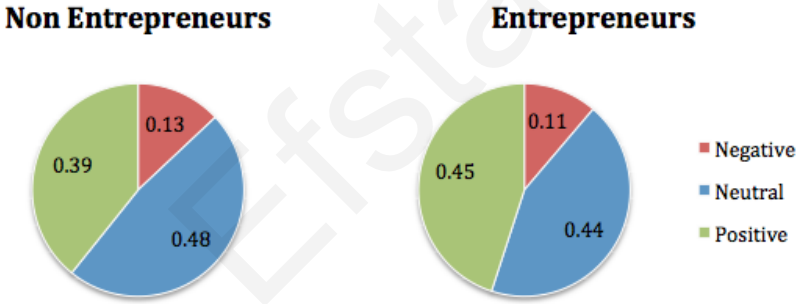
To verify this hypothesis we calculate and compare the average sentiment score of the tweets for the two categories over all the examined datasets. As described in section 5.2.1, sentiment scores range between -1 and 1. A score of less than 0 represents a negative sentiment, while one above 0 corresponds to positive sentiment. Tweets with a sentiment score of 0 are classified as neutral. Figure 5.1 plots the overall percentage of positive, negative and neutral tweets both for entrepreneurs and non-entrepreneurs for the three different geographical regions examined. As we can observe, the Twitter streams of entrepreneurs contain tweets that are significantly more positive than the tweets of non-entrepreneurs across all datasets ($p<0.005$).



(a) Los Angeles, USA



(b) London, UK

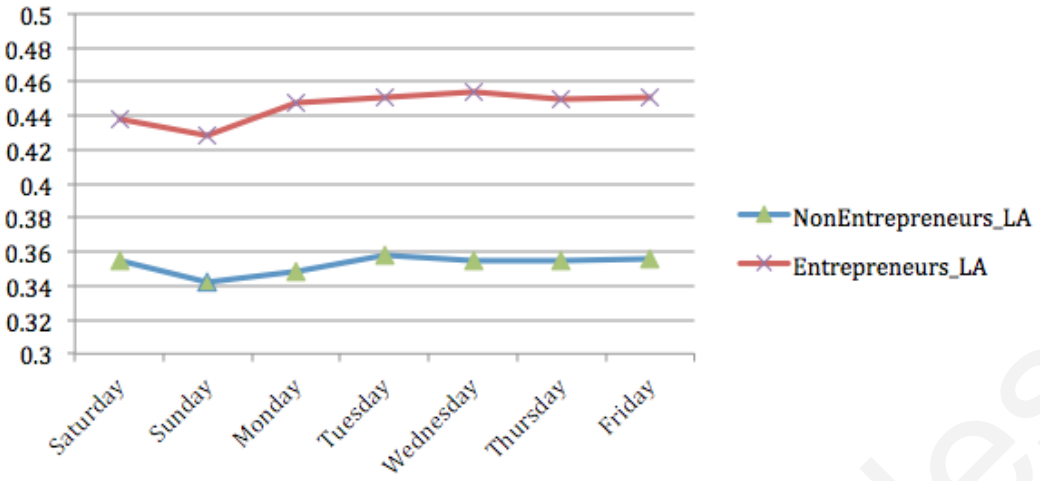


(c) World Wide

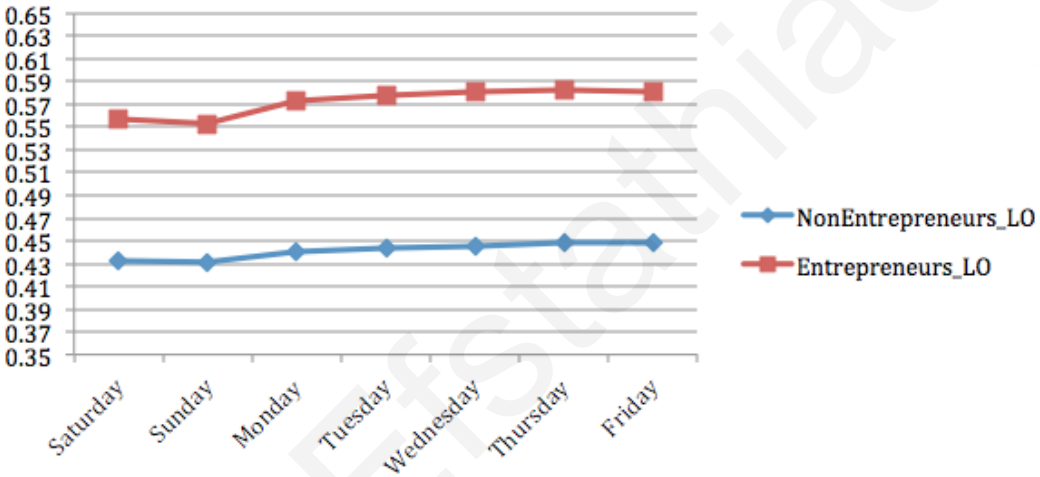
Figure 5.1: Tweet sentiment comparison between entrepreneurs and non-entrepreneurs.

The difference between the overall sentiment motivates us to further investigate the sentiment patterns of our two categories of interest. Figure 5.2 plots the percentage of positive tweets published per day of the week, for the two categories. As we can observe, for both user categories, the percentage of positive tweets is lower during weekends, while it increases during weekdays, reaching highest values towards the end of the week. In all cases, entrepreneurs are consistently more positive than non-entrepreneurs for each day of the week.

Sentiment is commonly used to measure the emotions of users' natural language. It can show the reaction of users to several events or their emotional state during a conversation. Combined with other information it can also show how different



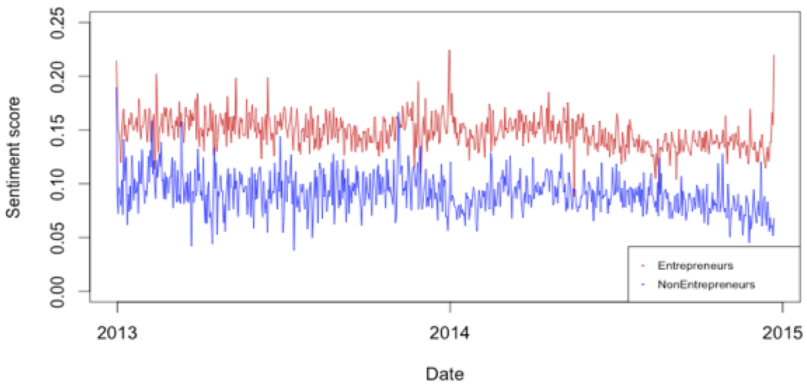
(a) Los Angeles, USA



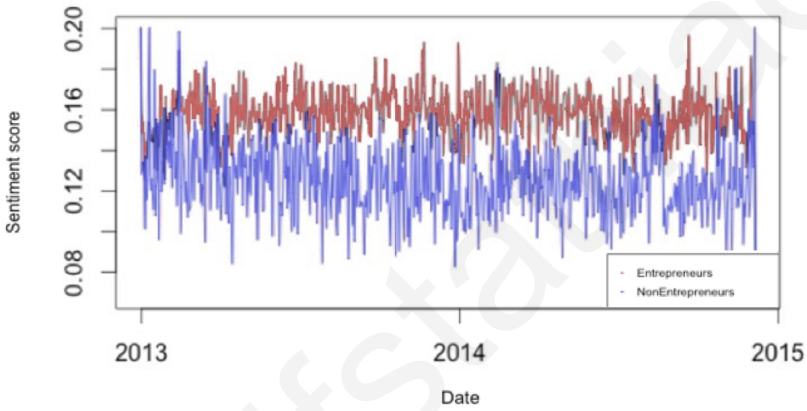
(b) London, UK

Figure 5.2: Comparison between entrepreneurs and non-entrepreneurs on their tweets; sentiment per weekday. We plot the percentage of positive tweets published from each category over the total number of tweets published by the same category during the specific day.

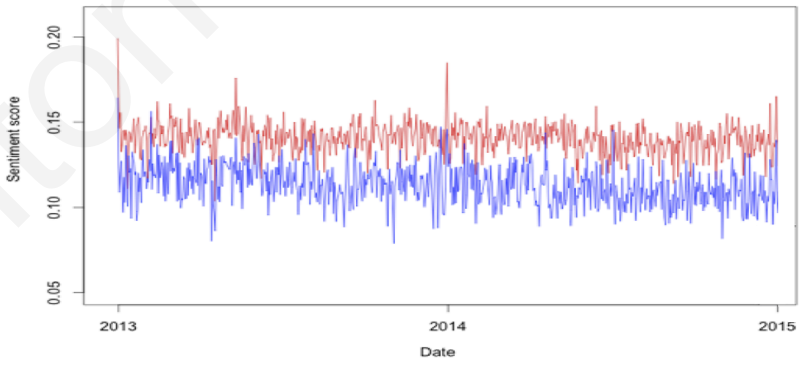
users react to specific events or express themselves during their everyday online interactions. As we observed so far, entrepreneurs and non-entrepreneurs differ in regard to the overall sentiment of their tweets.



(a) Los Angeles, USA



(b) London, UK



(c) WorldWide

Figure 5.3: Tweet sentiment comparison between entrepreneurs and non-entrepreneurs per calendar day, from January 2013 to January 2015.

Studying in-depth everyday conversations, we analyze the sentiment score of the two groups during extended periods of time. Figure 5.3 plots the sentiment

score of each calendar day during a period of 2 years; from January 2013 to January 2015. As we observe, "Happy" and "Sad" days can be identified in all timelines. Entrepreneurs have a sentiment score that is consistently more positive than that of non-entrepreneurs, by more than 14% on average, for the majority of calendar days between January 2013 and January 2015. Furthermore, non-entrepreneurs demonstrate more variance in their sentiment score for different calendar days. entrepreneurship field, which are not of high interest by non-entrepreneurs.

To further examine the validity of our hypothesis we proceed with the statistical analysis of our dataset through a fixed-effect regression function. In this function we compare a number of arguments with the possibility (expressed as the regression function dependent variable) of a tweet sentiment to be positive, neutral or negative.

The attributes used as inputs in the regression function are the following:

- Followers_count: This attribute indicates the number of followers a user has, meaning the number of other users who follow the specific one
- Followings_count: Indicates the number of followings a user has; the number of other users that he follows
- Statuses_count: This attribute indicates the number of tweets a user has published
- Entrepreneur: Is a flag that indicates if the specific user is an Entrepreneurs, according to the definition described in Section 3. If the user is an entrepreneur or non, this attribute has value 1 or 0 respectively
- isRetweet: Indicates if the specific tweet is a retweet of not, with values 1 and 0 respectively
- geotagged: Indicates if the specific tweet is geotagged or not, with values 1 and 0 respectively. If the tweet is geotagged then it has the specific latitude and longitude of the location that has been published from.
- Hashtag count: Is the number of hashtags the specific tweet contains.
- Followers_followee: Is a calculated attribute. For the calculation we use the result of number of followers divided by the number of followees he has.

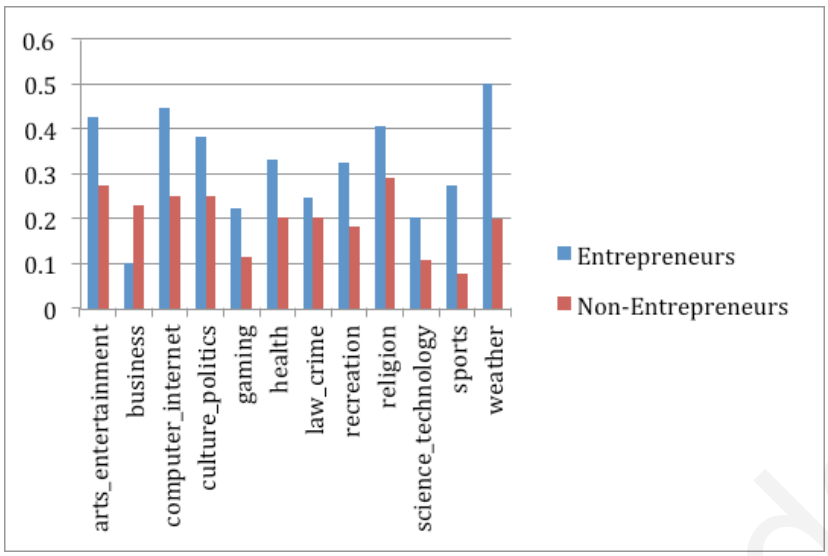
- Source: Indicates the application that was used for the generation and publication of the tweet. If a user published the tweet using his Android smartphone, this variable was coded as 1 and zero otherwise (e.g. iPhone, web).

Our results are shown in Table 5.5, columns (1)-(3), for the three different datasets, respectively. In all cases the attribute *Entrepreneur*, which denotes that the user is an entrepreneur, shows a strong correlation with the increased probability of a positive sentiment tweet. The *Entrepreneur* attribute is also the only attribute that shows correlation with sentiment consistently over all the examined datasets.

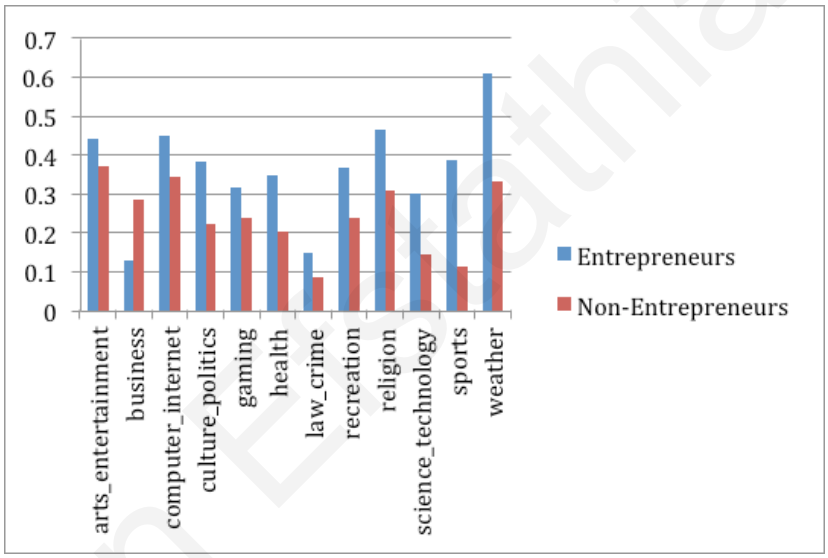
H2: Entrepreneurs are less likely than non-entrepreneurs to exhibit a positive sentiment in respect of business matters.

In order to examine the influence of the different topics on tweets' sentiment, we perform a study on the sentiment of these categories. For that purpose we retrieve tweets published from entrepreneurs and non-entrepreneurs in our datasets during a randomly chosen 2-month period of September and October of 2014. For the classification of tweets into different categories we use Alchemy Language API. Among other tools, Alchemy Language API provides the functionality of classifying text in specific categories, using natural language processing algorithms. Using this service we have classified 302,862 and 261,018 tweets in total, posted by entrepreneurs and non-entrepreneurs respectively.

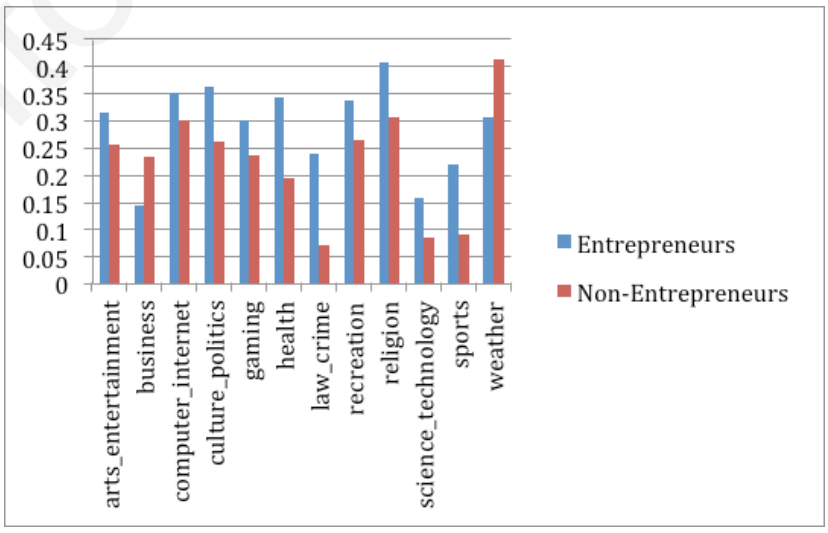
Figure 5.4, plots the sentiment score per concept for both entrepreneurs and non-entrepreneurs. As we can see, for all datasets, entrepreneurs are more positive than non-entrepreneurs in all topics except "business". We also performed statistical significance tests (T-tests) for all categories and the results show that there are significant differences in the sentiment across the concepts ($p<0.005$). Furthermore, we perform regression on non-business related concepts (Table 5.6), and show that entrepreneurs are more positive. This analysis step verifies our second hypothesis; entrepreneurs are less likely than non-entrepreneurs to exhibit a positive sentiment in respect of business matters. There is one minor exception – worldwide entrepreneurs feel more negative about the weather than non-entrepreneurs.



(a) Los Angeles, USA

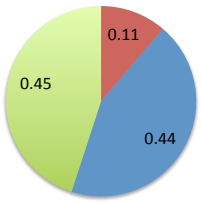


(b) London, UK

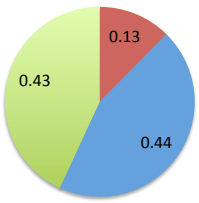


(c) World Wide

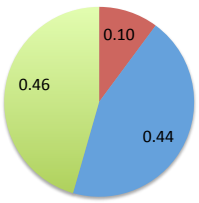
Figure 5.4: Sentiment score per concept, for a sample of Tweets published during 01/09/2014 - 31/10/2014.



(a) General-Traditional



(b) Serial



(c) Social

Figure 5.5: Overall sentiment for General-Traditional, Serial and Social Entrepreneurs. Positive, Negative, Neutral

H3(4): Social (Serial) entrepreneurs are more (less) likely than other entrepreneurs to exhibit a positive general sentiment.

Our results so far conclude that there is a significant difference between entrepreneurs and non-entrepreneurs for the three different datasets we have used. In this part we proceed in grouping entrepreneurs in sub-groups according to their entrepreneurship status. These two sub-groups have been identified based on the descriptions in their Twitter profile and correspond to "social entrepreneurs" and "serial entrepreneurs". We then analyze and compare the sentiment of the two, in order to investigate our H3 and H4 hypotheses.

Table 5.5, columns (5), reports our results for serial entrepreneurs, examining hypothesis H4. The coefficient on the serial variable, namely if an entrepreneur is described as "serial," is significantly negative, indicating that serial entrepreneurs experience lower average sentiment than all other entrepreneurs. Thus, we find support for hypothesis H4.

Table 5.5, columns (4), reports our results for social entrepreneurs, examining hypothesis H3. The coefficient on the social variable is significantly negative, indicating that social entrepreneurs experience higher average sentiment than all other entrepreneurs. Thus, we find support for hypothesis H3, as shown also on Figure 5.5.

5.4 Validation Tests

Our results so far suggest that entrepreneurs tend to tweet more positive content than non-entrepreneurs. In order to strengthen our analysis insights, we perform a series of validation tests, examining whether this positivity is correlated with the

fact that someone is an entrepreneur, or if it is influenced by other factors as well.

5.4.1 Terminology

As we observed earlier, entrepreneurs tend to tweet more positive content than non-entrepreneurs in the majority of the different concepts. However, due to the nature of Twitter, one could argue that people belonging to different categories may be tweeting about different issues or sub-topics under the same concept, and that this difference may affect the observed difference in sentiment between entrepreneurs and non-entrepreneurs. To explore this, we performed a group of verification tests, in order to strengthen the observation that entrepreneurs are more positive than non-entrepreneurs.

Initially, we randomly chose two concepts for which we conducted an in-depth terminology analysis. In particular, we selected all the tweets from entrepreneurs and non-entrepreneurs categorized in the concepts of "Religion" and "Health", tokenized them, removed stop words (namely, commonly used terms with no significant semantic weight, such as "the", "and", "or") and compared the remaining content using different methods. The first method that we use relies on monograms comparison: we take the terms that have been used in the tweets of the specific concept by entrepreneurs, and compare them with the corresponding terms used by non-entrepreneurs. The results show that the two groups use similar terms when talking about "Religion" and "Health", with the similarity on the top-100 most frequent terms reaching 81% and 73% for the London dataset, and 72% and 57% for the Los Angeles dataset, respectively.

Then, we proceed in n-grams analysis: with this methodology we compare the similarity regarding the position of the words in a sentence, and not only the appearance of a term. We take again the tweets that lie in the concepts of "Religion" and "Health", and proceed in applying n-grams comparison. Again, the results suggest that entrepreneurs and non-entrepreneurs of our dataset, use very similar terminology: for the concepts of "Religion" and "Health" the n-gram similarity reaches 69% and 61% for the London dataset, and 62% and 52% for the Los Angeles dataset, respectively. These similarity metrics provide a good indication that there is no significant difference in what entrepreneurs and non-entrepreneurs discuss over Twitter that would affect the sentiment score of the two groups' tweets.

Data	*Entrepreneurs, managers, directors, and executives*
Entrepreneur	0.013 *
	0.006
Followers	-0.13
	0.09
Following	9.10 ***
	1.85
Retweet	0.0017
	0.0012
Geotagged	-0.0016
	0.0010
Hashtag count	0.0020
	0.0014
#tweets	-0.00000042
	0.00000027
Constant	0.12 ***
	0.00
Sample	Worldwide
F-Stat	4.532 on 6 and 314 DF
p-value	0.0002004 ****

*Table 5.7: Sentiment of Entrepreneurs compared with Managers, Directors, and Executives. ($p < .0001$ '*****', $p < .001$ '****', $p < .01$ '***', $p < .05$ '**', $p < .1$ '*')*

5.4.2 Entrepreneurs Vs Managerial Positions

Using the profile description field, we identified 199 non-entrepreneurs in the Wold-Wide dataset, who hold a managerial positions, as denoted by profile terms of "director", "manager", "executive" etc. We compared the sentiment of their tweets against that of entrepreneurs in the same dataset. Our results suggest, again, that entrepreneurs are more positive than executives. Table 5.7 presents the fixed-effect regression statistics.

5.4.3 Profile Description

We define as an Entrepreneur any Twitter user who uses at least one of the following terms in their profile description field: entrepreneur, founder, co-founder, start-up, start up, business-owner, business owner. For validation reasons, from this group of users, we filter out the subsample that uses only the term of 'co-founder' and compare their sentiment with non-entrepreneurs. The results are similar with the general insights of our study, having entrepreneurs being overall more positive.

Data	*All*	*All*	*All*	*Entrepreneurs only*	*Entrepreneurs only*
Region	London	Los Angeles	Worldwide	Worldwide	Worldwide
	(1)	(2)	(3)	(4)	(5)
Entrepreneur	0.081 ***	0.088 ***	4.465e-02***		
	0.000	0.00	0.002		
Social entrepreneur				0.024***	
				0.004	
Serial entrepreneur					-7.8E-9 *
					3.8E-9
Followers	1.72	0.39 ***	9.283e-10	-0.062 *	-0.079 *
	2.54	0.07	0.12	1.172e-08	0.038
Following	9.52	-0.014	8.641e-07***	9.75 ***	9.78 ***
	6.59	0.17	2.33	2.328e-07	0.71
Source	-1.4E-7	-1.4E-7			
	2.0E-7	1.1E-7			
Retweet	-0.028	-1.4E-6	-1.748e-03	5.7E-8	5.1E-8
	0.017	1.9E-6	1.805e-03	5.4E-8	5.1E-8
Geotagged	-0.0032	-0.027	-2.411e-03	0.00068	-0.027
	0.056	0.019	1.804e-03	0.00186	0.019
Hashtag count	0.0048	-1.2E-7	-3.147e-03	0.00022	0.0011
	0.056	-1.0E-7	1.804e-03	0.00019	0.0024
Followers followee	5.01E-06	-1.38E-06	5.2E-8	1.6E-6	1.7E-6
	4.57E-06	1.85E-06	6E-7	9E-7	1.4E-7
#tweets	-2.57E-04	7.06E-09	-4.58e-07	-3.22E-07**	-1.23E-07
	4.97E-05	1.52E-07	2.7E-7	6.31E-08	1.01E-07
Constant	0.18 **	0.11 ***	0.097 ***	1.39E-01 ***	1.63E-01 ***
	0.057	0.00	0.003	9.24E-04	1.66E-02
N	3000	8124	50360	25180	25180
F-Statistic	69.88	1.976	80.86	2.634	15.45
P-Value	$< 2.2e-16$ * **	0.05444 * *	$< 2.2e-16$ * **	0.003128***	$< 2.2e-16$***

*Table 5.5: Regression results for all datasets, having as dependent variable the sentiment score. ($p <$ .0001 '*****', $p <$ .001 '****', $p <$ .01 '***', $p <$ .05 '**', $p <$ .1 '*')*

All	*All*	*All*	*All*
Region	London	Los Angeles	Worldwide
Entrepreneur	5.084e-02***	5.091e-02***	1.173e-01***
	1.215e-02	1.041e-02	1.659e-02
Followers	8.433e-06	2.510e-05***	4.810e-06
	6.458e-06	5.856e-06	5.448e-06
Followings	1.065e-05	-3.052e-05**	8.019e-06
	9.249e-06	9.552e-06	1.052e-05
follower followee	-2.609e-04	-6.838e-04***	7.523e-05
	2.001e-04	1.706e-04	2.373e-04
Retweet	2.880e-04	-4.122e-02	-7.002e-02
	9.635e-03	1.959e-02	1.156e-01
Geotagged	-8.361e-03	-9.576e-03	-7.084e-03
	9.634e-03	1.979e-02	1.128e-01
Hashtag count	1.566e-02	-4.601e-02*	1.727e-01
	9.633e-03	1.959e-02	1.244e-01
F-Statistic	7.593	10.14	12.27
P-Value	3.553e-09***	9.851e-13***	1.043e-15***

*Table 5.6: Regression results for all non-business tweets, having as dependent variable the sentiment score. ($p < .0001$ '*****', $p < .001$ '****', $p < .01$ '***', $p < .05$ '**', $p < .1$ '*')*

Chapter 6

Online Social Networks Evolution: Revisiting Twitter Network

In their popular study of the Twitter network, Kwak et al., examined the full Twitter graph as it appeared in 2009 [58]. The dataset that they have collected and studied is the largest publicly available Twitter dataset according to the number of nodes and edges. With their analysis they provided insights about the overall network topology, online activity of the users and influential users that existed at that time. Their results summarize the characteristics of Twitter in 2009 and its power as a new medium of information sharing. With this study we revisit the same sample of users and collect the full information that is available from the Twitter API. We collect a total of 34.6 million user profiles, connected through 2.05 billion relationships. Based on the provided insights and data, we aim in analyzing the Twitter network as is today, and provide a comparison with the snapshot of 2009.

We address the different characteristics of the 2009 Twitter network, as it appears to be connected today, and examine the changes in connectivity of the network in general and the users in particular. To the best of our knowledge this work is the first quantitative study on the entire Twittersphere, that examines the long term evolution of the Twitter network.

Our contributions can be summarized as follows:

1. We observe a network that gets denser through the years, with the number of edges between the users in 2015 being almost double than 2009.

2. We clearly observe a "rich-get-richer" phenomenon, since the increased number of edges is mainly directed towards the most popular users.

Snapshot	*Vertices*	*Edges*	*Density*
TW2009	40,103,281	1,468,365,182	1.83e-6
TW2015	34,664,106	2,056,655,361	3.42e-6
TW2009C	34,664,106	933,256,652	1.55e-6

Table 6.1: Description of the 3 different Twitter graph snapshots.

3. Despite the increased number of edges, network connectivity seems to be decreasing. The Largest Strongly Connected component of the network decreases by 20%, in number of nodes, showing that the connections not only increase in total but are also redirected.

4. In the 2009 most of the popular users where popular in both followers and PageRank classification. Our study reveals a decoupling of the two methods, where most popular users through PageRank are not necessarily the ones with the highest in-degree.

5. We identify the reasoning behind users who left the Twittersphere and correlate it with their position in the graph. Our analysis suggests that users who have been banned from Twitter have different degree distributions than others, while the participation in the largest Strongly Connected Component of users who intentionally left the network is by 10% higher than the rest. Furthermore, PageRank classification suggests that several users maintained highly ranked positions before their disappearance.

The remainder of this chapter is structured as follows: We describe the experimental setting and the datasets used in the study in Section 6.1; Section 6.2 presents the topological analysis of the Twitter network and the comparison between the different snapshots. In Section 6.3 we rank users based on the number of followers and PageRank, and compare the results with the ones of Kwak et al. study [58]. Finally, Section 6.4 describes the study performed on users who have been disappeared from the network and present the derived insights.

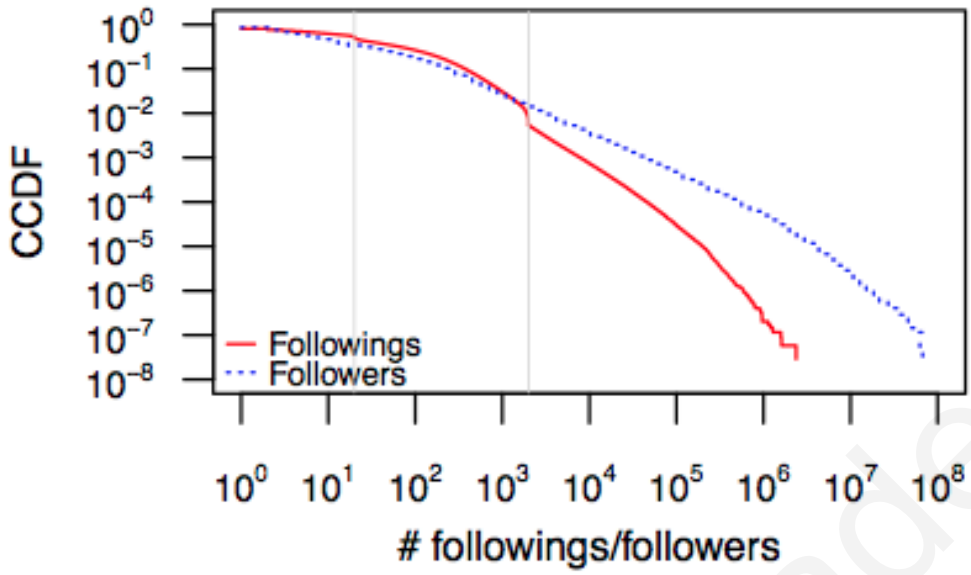


Figure 6.1: Complementary Cumulative Distribution Function (CCDF) of followings and followers.

6.1 Collected Data

Our analysis is based on two different snapshots of the same Twitter network: (*i*) the complete Twitter 2009 graph, as collected and shared by [58], and (*ii*) the collection of the same list of Twitter users and their social graph as it appeared in late 2015. The 2009 graph was made available by Kwak et al.[1] According to the authors, the dataset represents the complete social graph of Twitter in 2009. Using the list of Twitter users that appeared in *TW2009* we perform a large-scale collection, through the current version of the Twitter API[2], with respect to platform's terms of use and users' privacy.

In order to collect this large scale Twitter dataset in a short-period of time we perform a distributed data collection campaign. Since Twitter API policy has been updated from IP-based to Application-based [57], we follow a crowd-crawling approach asking Twitter users to authorize our multiple applications to make request for public information on their behalf. We manage to configure a large number of Twitter applications instances in order to reduce the waiting time between the requests[3]. We implement this approach on 3 different machines; an action that enables us to collect the ego-networks of 1.2M users per day. [28]

[1] `http://an.kaist.ac.kr/traces/WWW2010.html` (Last accessed: Jun. 2016)
[2] `https://dev.twitter.com/rest/public` (Last accessed: Jun. 2016)
[3] `https://dev.twitter.com/rest/public/rate-limiting` (Last accessed: Jun. 2016)

Through this collection we retrieve the same set of Twitter users and their ego-network state (followers and followings) in November 2015. From this network we remove any connections (edges) that are directed towards or coming from users who do not belong in 2009 set. Thus, our *TW2015* snapshot contains only the connections that existed and have arise between the users that consisted the Twitter social network in 2009.

Table 6.1 presents the details of the two snapshots. As a first general observation we can see that more than 5 million users from *TW2009* have disappeared in the *TW2015* snapshot. The reason for a user not to appear in the snapshot can be explained through three different scenarios, based on Twitter API response when requesting the specific data: (*i*) the user has been banned from the network due to violations of the terms of use (*ii*) the user intentionally removed her account deleting herself from the Twitter Online Social Network platform (*iii*) the user updated her privacy settings and made her information (profile and ego-network) private (not publicly accessible through the Twitter API). We further examine the properties of these three user categories in Section 6.4.

In addition to the two full graphs of the 2009 Twitter users we also examine and compare, where relevant, the 2009 graph as it would appear if the users that belong to the above three categories where not existent in 2009. The *TW2009C* presents the snapshot of this case.

As we can see from Table 6.1, the social network has become denser through the years. The connections between the same network users in 2015 are more than double the ones that existed in 2009 (*TW2015* Vs. *TW2009C*). This observation shows that the same set of users are constantly identifying each other creating new connections between them and becoming interested in the content they share. In the next section we examine in more detail the difference between the snapshot connections, trying to identify whether these new relationships are additional to the ones existed in 2009 or whether there is a general move of connections, with some users losing their followers while others gain more attention.

6.2 The Twitter Graph Evolution

In this section we present a study on the different snapshots of the 40.1M users of Twitter, regarding their graph metrics. For each analysis step we describe the proce-

dure followed, results and derived insights. Furthermore, we present a comparison between the networks and discuss their topological differences.

6.2.1 Basic Analysis

Similarly to Kwak et al. [58], we analyze the characteristics of the followers and followings of the TW2015 users. The following relationship is directly related with a user's action: the individual chooses to follow another profile due to her own reasoning. On the other hand, the follower relationship is influenced by an indirect action; an individual maintaining an active profile, posting interesting content, with the goal of and attracting new followers.

Figure 6.1 plots the complementary cumulative distribution function (CCDF) of the number of followers (dotted blue line) and followings (solid red line). The followings case presents similar glitches as in 2009. According to Kwak et al. [58] the glitch on $x = 20$ is an inherent consequence of the Twitter initial recommendation of 20 people to follow, when a user first creates an account. The observation of the same glitch in the network snapshot taken 6-years later, shown by the first vertical line in Figure 6.1, intrigue us to investigate this further. We analyze the group of users who follow exact 20 other accounts, and we see that on average they have less than 19 followers while their average number of tweets is less than 32. With this in mind, we conclude that these are inactive users who did not use Twitter after their initial sessions.

The next glitch that we observe is the one at around $x = 2000$. Twitter imposed a limit at 2,000 followings, for each user. After that number any increase in the number of followers is correlated with the follower-following ratio and it is account specific. Myers et al. [79] in 2014, examined this limit and concluded that the platform does not allow users to follow more than 2,000 accounts unless they themselves have more than 2,200 followers. This threshold has been updated to 5,000 followings, according to Twitter[4]. For an account to get 2,200 followers is a difficult process, needing a lot of effort to interest people. Most accounts will never be able attract that many followers and for this reason will remain in the limit imposed by the service. However, this does not mean that the set of followings of an account remains unchanged through the years. Users may select to remove accounts that are not interesting anymore, in

[4]https://support.twitter.com/articles/66885 (Last accessed: Jun. 2016)

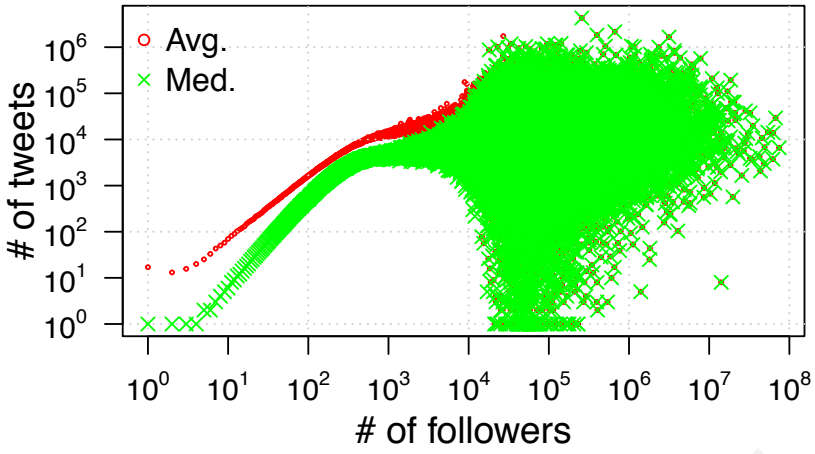


Figure 6.2: The number of followers and that of tweets per user.

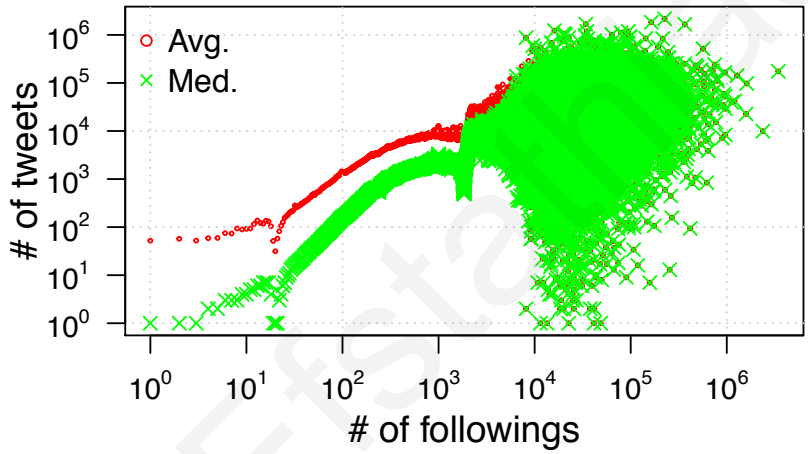


Figure 6.3: The number of followings and that of tweets per user.

favor of new more interesting accounts.

The follower distribution in Figure 6.1 shows the presence of several celebrity users in Twitter. These users attract more than 10K followers. The large majority of users has less followers and fits to a power-law distribution with an exponent of 2.101. This value is lower than the 2.276 observed by Kwak et al., however it remains in the boundaries between 2 and 3 that characterize the majority of real-world networks, including OSN [58].

6.2.2 Followers vs. Tweets

A common perception regarding Twitter is that the more active a user is (that is the more content she shares), the more followers (attention) she gets. Kwak et al. observed this perception to be true only for users with up to 5,000 followers. After that point there was no obvious correlation between the number of followers and

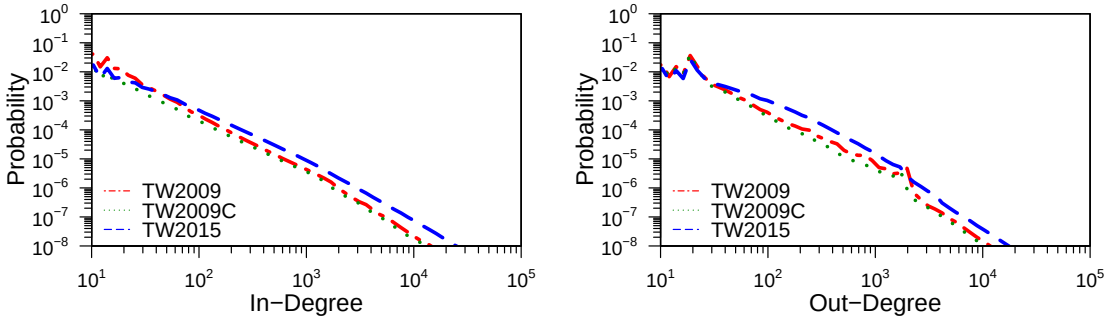
that of the tweets of a user. Figure 6.2 presents the results of the same analysis for the TW2015 snapshot. The figure plots the number of followers as a function of the median (green cross) and average (red circle) number of tweets for each user. Comparing with the TW2009 study, we can see that the results are similar for users who have less than 10 followers; the majority has never tweeted or did just once, maintaining a median value of 1. Similarly, the existence of outliers who tweeted much more than the expected, based on their followers counter, preserve an average value always higher than the median in regards to the number of tweets. Furthermore, the flat line observed between the values of 100 and 1000 followers in 2009, has been moved to 1000 and 1500, as the number of followers seems to be increasing.

Figure 6.3 examines the relationship between the activity of a user (number of tweet she posts) as the number of the people she follows increases. It plots the number of followings and median and average number of tweets per each user in our dataset. The two irregularities at $x = 20$ and $x = 2000$ observed in Figure 6.1 also appear in this plot. Furthermore, the additional irregularities observed by Kwak et al. and attributed to spam accounts have disappeared, an expected consequence since Twitter removed these accounts.

6.2.3 Degree Distribution

Twitter follower graph is a directed graph $G = (V, E)$, where each vertex $v \in V$ represents a user in the network while each edge $e \in E$ represents a directed follower relationship between the 2 vertices $\{v_s \rightarrow v_d\}$. Thus, each vertex has an in-degree, which represents the number of its followers, and an out-degree which represents the number of its followings.

In the previous section we observe that the number of connections between the Twitter users has almost doubled during the time period separating the two collections. In this section we examine the degree distribution of the three networks to answer whether this increase can be attributed to a small number of nodes that increased their incoming or outgoing connections tremendously or whether the majority of users has participated in this increase. Since Twitter is a directed network we examine both the in-degree, the number of the user's followers, and out-degree, the number of a user's followings.



(a) In-Degree

(b) Out-Degree

Figure 6.4: In-degree and Out-degree of the 3 different Twitter snapshots.

	25%	50%	75%	100%
In-TW2009	2	8	17	2.99M
Out-TW2009	4	9	21	770K
In-TW2009C	2	4	8	2.57M
Out-TW2009C	3	9	20	662K
In-TW2015	2	5	19	3.21M
Out-TW2015	6	20	69	608K

Table 6.2: Statistics of the average degree distributions for the 3 networks.

Table 6.2 shows the 25th, 50th, 75th and 100th percentiles of the degree distribution in the different Twitter snapshots. With an exception for the popular users (100th percentile), in all other cases the in-degree is smaller than out-degree. In Twitter terms, this means that the average user is being followed by less users that the ones she follows. This observation holds for both 2009 and 2015 graphs, which reveals a non-era binded finding. Additionally, we can observe a rich-get-richer phenomenon, as the in-degree of the most popular users increases. On the other hand, less popular users show an out-degree that almost triples in some cases. This observation leads us to conclude that the increase in the number of edges observed from TW2009 to TW2015 is due to popular users getting more follow relationships coming from the rest of the network.

Figure 6.4(a) plots the in-degree distribution for the 3 different snapshots. As expected, we observe a heavy tail power-law distribution in all cases. From Figure 6.4(b) we can see that the out-degree also follows a heavy tail power-law distribution but not at the same extent as in-degree; a fact also described by [79].

Figure 6.4(b) presents a spike at the value of 2000 out-degree nodes for both

2009 snapshots. We examine the reasoning behind it in Twitter mechanisms and found that the platform applies an anti-spam/bots strategy regarding the number of accounts that an individual user can follow. In recent studies spammers and bots have been characterized by very low values of *#Followers*/*#Followings* ratio [20]. For that reason Twitter sets a limitation of 2000 followings for each account who has less than 2200 followers [79]. In the 2015 snapshot we do not observe similar spike, as the threshold strategy has been changed during recent years[5].

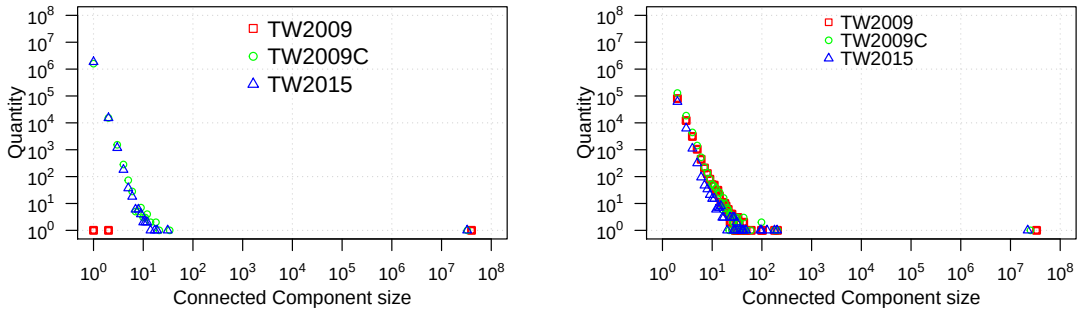
Discussion: The results suggest that the increase in the number of edges observed from TW2009 to TW2015 is due to the increasing number of connections that popular users attract, coming from the rest of the network. Establishing a relationship in Twitter denotes that a user follows another and is able to receive notifications and read the content the latter publishes. However, the large number of out-degree, and thus followings, would reasonably be a problem for a user to easily access and read information of interests. However, as reported in [58], Twitter looks like more an information network instead of a social network. Furthermore, the platform of Twitter provides to its users the functionality to 'Mute' an account; the following relationship remains but the content that the muted user publishes does not appear in the tweet feed of the one who muted her.

6.2.4 Connected Components

We now turn our attention in examining how the connectivity of the network changes, as this can be seen through the number of Strongly and Weakly Connected Components. A Strongly Connected Component in a directed graph is a subgraph where there exists a path from every node to every other node in the subgraph. A Weakly Connected component is a sub-graph where all nodes are connected with some path, ignoring the direction of the edges.

Figure 6.5 plots the distribution of the size of the Weakly and Strongly Connected Components for the 3 Twitter snapshots. As we can see from Figure 6.5(a), in all cases a large connected component maintains an enormous size compared to the rest components. In Twitter 2009 graph a single Weakly Connected component covers more than 99.9% of the nodes. Despite the fact that the number of Weakly Connected Components has been increased in the graph of 2015, the coverage of the

[5] `https://support.twitter.com/articles/66885` (Last accessed: Jun. 2016)



(a) Weakly Connected Components (b) Strongly Connected Components

Figure 6.5: Strongly and Weakly Connected Components of the 3 different Twitter snapshots.

largest WCC is still very high, as it contains 94.58% of the graph nodes. We also see that 5.52% of the nodes change Weakly Connected Components in 2015 snapshot. Finally, if we exclude removed users from 2009 graph, we observe an increase on the quantity of WCC, while the largest contains 95.58% of the nodes.

Studying the Strongly Connected Components enable us to extract more interesting insights for the case of Twitter, as in such components the direction of the edge is not ignored. Due to the fact that Twitter graph is directed the metric has different meaning than WCC. From Figure 6.5(b) we observe that in all 3 cases the largest Strongly Connected Component covers the largest portion of the graph, while several others maintain a much smaller size. In 2009 the largest SCC covered a large percentage of the graph, 83.90%, a higher value compared to the 65.56% of the largest SCC in 2015. The largest SCC in 2009 seems to be closer to the coverage observed in other social graphs, such the ones of MSN messenger [61] and Facebook [99], which show a coverage of more than 99%.

An important observation from both cases is that despite the fact that the network is becoming denser, it seems to be disconnecting. While the largest WCC in 2009 included almost all the network, in 2015 the number of WCC increases, showing a number of sub-graphs that are disconnected from the rest of the network. Furthermore, the largest SCC size decreases by almost 20%, and we observe an increased number of smaller SCC. Taking into account that popular users are the ones that actually increase their incoming connections, we might consider that Twitter users decide to remove edges from non-popular users to target more popular ones. This change limits the paths that connect users between them, resulting in more groups of fewer nodes that can be reached by each other.

To put this into perspective, we calculate the percentage of users who appear in different Connected Components in 2009 and 2015 snapshots. The comparison regarding Weakly Connected Components show that 5.52% of the nodes change component in the evolving snapshot, while none of them left the largest WCC. In contrast, in the case of the Strongly Connected Components we observe that 72.43% of the vertices have moved to a different one during the 6 years period, having 22.94% leaving and 4.60% joining the largest.

Discussion: From the results we derive the insight that the structure of the network has changed significantly regarding the Strongly Connected Component. We observe a decrease of about 20% in the coverage of the largest SCC between 2009 and 2015 snapshots. One possible reasoning of this fact is that Twitter in its early years was used as a social networking platform. As the years past, the network has been evolved and changed to a more information dissemination platform, where users connect with accounts who post content that lies in their interests instead of the one with whom they share physical-world relationship. This resulted to a more sparse network, where clusters between users who share similar interests have been created.

6.2.5 Reciprocity

As presented in [58], the reciprocity of the 2009 Twitter snapshot does not exceed 22.1%, meaning that only this amount of user pairs follow each other. The rest 77.9% of the pairs are single sourced, thus they only share one relationship. The reciprocity in 2015 Twitter snapshot increases to 29.2%, showing that more and more users tend to follow back the ones who follow them. However, this number is still much smaller than the reported values of 68%, 79% and 84% for Flickr [15], Yahoo!360 [55] and YouTube [74] respectively.

Reciprocity in directed Online Social Networks is often considered as a measure of a stronger connection between two users [2]. However, in Twitter the follow-back mechanism is also used for more practical reasons, such as to recruit more followers. A number of users mentions in their description fields that they follow back, in order to attract others to follow them so as to increase their number of followers. As we observe, that mechanism in not popular as only 10,656 users have added this information in their profile description fields.

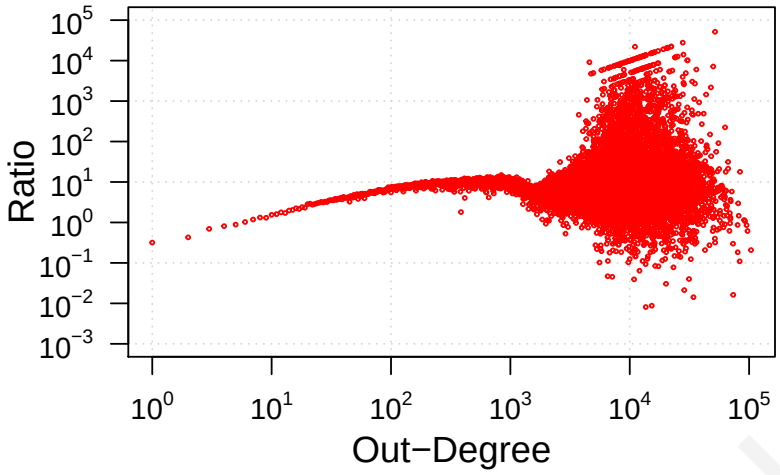


Figure 6.6: The Out-Degree and the ratio between newly created and removed out-going edges.

6.2.6 Edges Comparison

The study of the users behavior in OSN platforms has gain the attention of researchers during past years. The majority of the studies has been focused on the content that individuals publish and how their behavior change in time [49, 56, 66]. Having collected two different snapshots with a difference of several years, we study the behavior of users regarding the connections created and removed over the time.

Figure 6.6 plots the average value of the ratio between newly created and removed out-going connections in relation to the number of out edges of each node (out-degree). Users with an out-degree of less than 5,000 tend to remove 1 edge for every 6.33 created. For users with an out-degree of less than 100 this ratio decreases to 3 on average. The latter result is similar with the one estimated by Myers et al., who observe 1 removal for every 3 created [78]. Furthermore, for users who have an out-degree of more than 500, the ratio between newly created and removed edges is 5.73. The results suggest that users tend to create edges in a higher rate than removing. However, this ratio does not exceed the value of 8 for any case when $x <= 10,000$.

With Figure 6.7 we examine the average value of the ratio between newly created and removed in-coming connections in relation to the number of in-coming edges of each node (in-degree). As we can see, the ratio is increasing steadily for users with less than 100 followers until it reaches a value of 3.21, while it remains almost stable at 3.51 between $x = 100$ and $x = 1000$. Users who maintain an in-degree between 1,000 and 5,000 edges tend to lose 1 follower for every 4.24 new in-coming

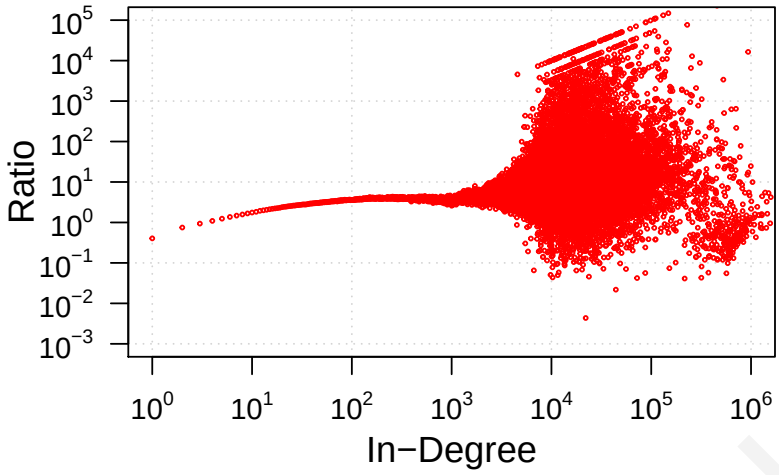
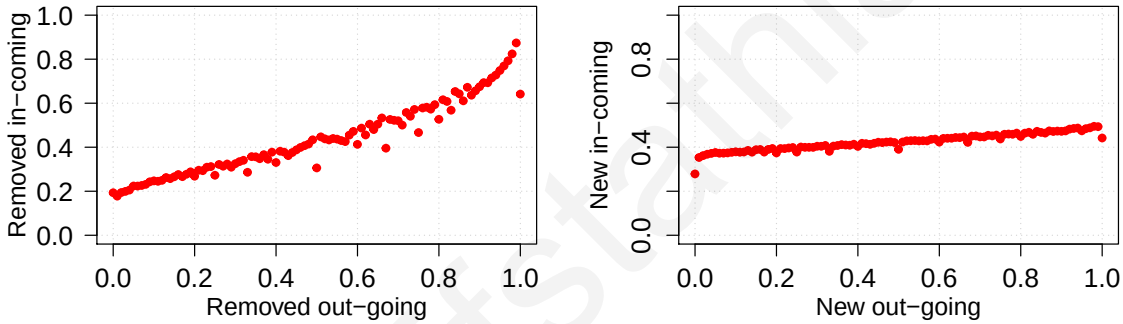


Figure 6.7: The In-Degree and the ratio between newly created and removed in-coming edges.



(a) Fraction of removed out-going and in-coming edges.

(b) Fraction of newly created out-going and in-coming edges.

Figure 6.8: Fractions of removed and newly created edges.

connections, while for more popular users, between 5,000 and 10,000 followers, the ratio increases to 5.44. Furthermore, for users who maintain an in-degree of more than 10,000 edges, which are mostly celebrities, the ratio between newly created and removed edges decreases to 3.66.

Figure 6.8(a) plots the fraction of removed out-going connections as a function of the average fraction of removed in-coming connections. As we can derive, the fraction of out-going connections removal increases linearly with the one of the in-coming connections. From this result we can conclude that users un-follow other accounts in similar rate as their followers remove the connections towards them. In Figure 6.8(b) we plot the fraction of newly created out-going connections as a function of the average fraction of newly created in-coming connections. As we can see, there is a slight increase of new in-coming connections related with the increase

of new out-going connections. However, this increase is not at the same scale as the one in out-going connections.

Discussion: These results enrich the hypothesis that in-coming connections are not directly related with a user's action; a user cannot increase her in-coming relations by direct actions, unlike the out-going relations. In order to attract more followers a user should post content that fits the interests of others, and trigger them to follow her account, instead of following other to follow her back.

6.2.7 Degree of Separation

The small world phenomenon refers to the surprisingly small distance that actually separates two users in a social network. Kwak et al. [58], examined the full Twitter graph as it appeared in 2009 and their analysis on the structural properties of the graph shows an average shortest path length of 4.12.

The calculation of the average shortest path between all pairs of vertices is computationally infeasible, due to the large scale of the collected dataset. Thus, we employ a sampling procedure similar to the one performed by [58]; we randomly retrieve a group of 2,000 users, that we call *seeders* and calculate the shortest path between them and all other vertices in the graph. The calculations have been performed using the single source shortest path algorithm, which we have developed on GraphChi [60]. Figure 6.9 presents the results on the degrees of separation between the seeders and all other vertices in the network. Our results show that the distance between two nodes in the graph is 4.05, while the median is 4.29 intermediates. As we can observe, the average shortest path value has been slightly decreased in the past 6 years.

Discussion: Despite the fact that a large number of Twitter users have been disappeared from the network, as presented in Section 6.4, the length of the average shortest path has been reduced. At the time of Kwak's et al. study [58], Twitter graph had an average value much smaller than Facebook; users where separated by 4.12 and 4.74 intermediaries on average respectively. However, a recent study from Facebook[6] shows that the average degree of separation in the network gets smaller

[6]Three and a half degrees of separation, `https://research.facebook.com/blog/three-and-a-half-degrees-of-separation` (Last accessed: Jun. 2016)

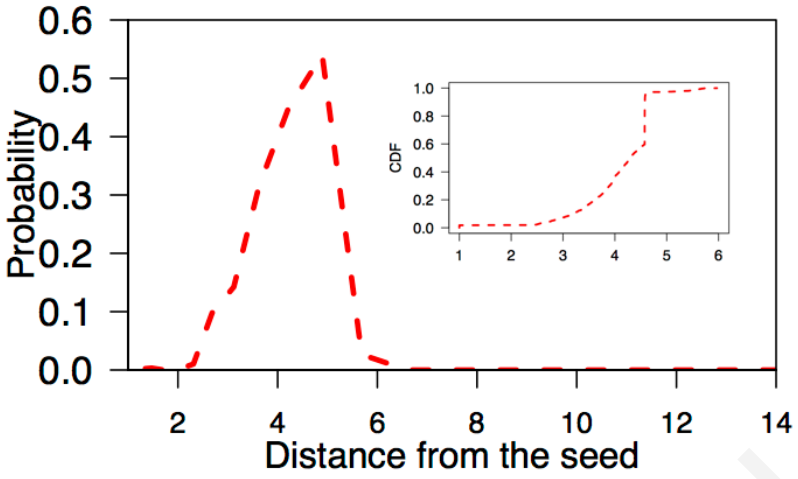


Figure 6.9: Distribution of degrees of separation between 1000 random chosen users and the rest of the network. Inner plot shows the cumulative distribution function for the same shortest paths.

and reaches an average value of 3.57 intermediaries while within the US, people are connected to each other by an average of 3.46 nodes. Compared to our observation of 4.05, we conclude that the average shortest path in Twitter decreases in time but with much smaller coefficient than Facebook. This result could be explained by the different type of the two graphs, as Twitter is directed while Facebook is undirected.

6.3 Rankings

In this section we present the results regarding two different popularity metrics on Twitter users. For each profile, Twitter maintains a counter that reveals the number of users who follow the corresponding profile. As reported by Kwak et al. [58], this metric does not reflect the topological influence of the node; i.e. the number of influential users who follow her. Thus, we proceed to another ranking procedure, using the widely used PageRank algorithm on the collected social graph [85].

6.3.1 By Followers

We use the straightforward approach of ranking users by descending order based on the number of their followers. As shown on Table 6.3, the users contained in this list are very different than the one published in 2009 [58], as we observe 65% new entries. From the rest 7 users, only Barack Obama manage to improve his corresponding 2009 position, while Oprah Winfrey and CNN Breaking News accounts, who appeared

	Ranking by Followers			Ranking by PageRank		
Rank	Screen Name	Name	Change	Screen Name	Name	Change
1	katyperry	KATY PERRY	New	TheEllenShow	Ellen DeGeneres	+3
2	justinbieber	Justin Bieber	New	BarackObama	Barack Obama	=
3	BarackObama	Barack Obama	+4	cnnbrk	CNN Breaking News	=
4	taylorswift13	Taylor Swift	New	twitter	Twitter	+5
5	YouTube	YouTube	New	aplusk	ashton kutcher	-4
6	ladygaga	Lady Gaga	New	britneyspears	Britney Spears	-1
7	jtimberlake	Justin Timberlake	New	Oprah	Oprah Winfrey	-1
8	TheEllenShow	Ellen DeGeneres	-5	jimmyfallon	jimmy fallon	+4
9	twitter	Twitter	-3	nytimes	The New York Times	+7
10	britneyspears	Britney Spears	-8	KimKardashian	Kim Kardashian West	+10
11	KimKardashian	Kim Kardashian West	-1	RyanSeacrest	Ryan Seacrest	-1
12	shakira	Shakira	New	TheOnion	The Onion	+7
13	selenagomez	Selena Gomez	New	SHAQ	SHAQ	-6
14	ArianaGrande	Ariana Grande	New	lancearmstrong	Lance Armstrong	-3
15	ddlovato	Demi Lovato	New	taylorswift13	Taylor Swift	New
16	Oprah	Oprah Winfrey	-11	StephenAtHome	Stephen Colbert	New
17	cnnbrk	CNN Breaking News	-13	stephenfry	Stephen Fry	+1
18	jimmyfallon	jimmy fallon	New	mashable	Mashable	New
19	Pink	P!nk	New	google	Google	New
20	Drake	Drizzy	New	justdemi	Demi Moore	-6

Table 6.3: Top-20 users ranked by the number of followers and PageRank in the Twitter 2015 social graph. Users who belong in both lists are highlighted. Column Change reports the update from TW2009 position in Top-20 rankings.

in top-5, are now outside of the top-15 rankings.

6.3.2 By PageRank

We apply the PageRank algorithm on Twitter 2015 graph, which contains 34.6M users, connected with 2.05B directed edges. Each node of this network represents a user, while each edge a following relationship. We calculate the PageRank value using the GraphChi cpp implementation. Table 6.3 shows the top-20 list regarding this value, with a column that describes the updates regarding their difference between the 2015 and 2009 rankings. As we can see, despite the fact that users of 2015 list are by 80% the same, only 10% of them maintain the same rank position. Furthermore, we observe that 35% of the top-20 entries have improved their rankings, while the same fraction appears in a position lower than in 2009.

6.3.3 Discussion

From the comparison between the 2009 and 2015 top-20 rankings lists we observe significant differences. We find this differences related with physical world events, i.e Barrack Obama maintains or improves his rankings as he also upholds his influence in the physical world. On the other hand, Ashton Kutcher appeared as 1st in both Followers and PageRank 2009 rankings before his famous divorce with Demi Moore; for the 2015 rankings he is outside top-20 and in 5th position, respectively.

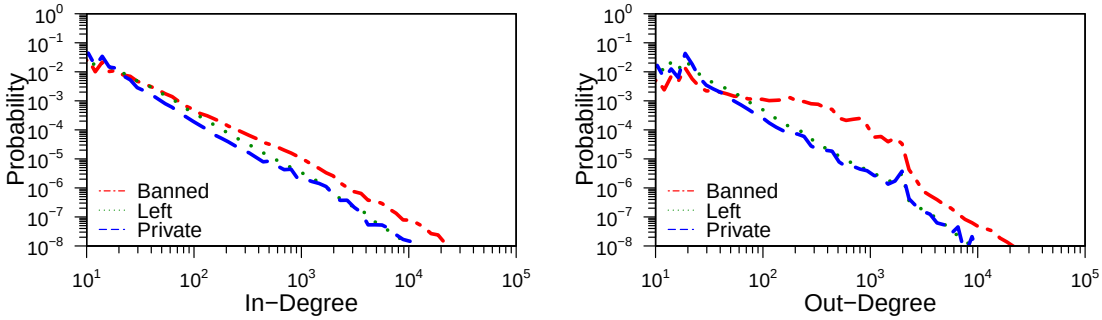
Comparing the two lists of 2015 we can see that only less than half of the users is presented in both. As we observe, the top-2 users in followers rankings do not belong in the PageRank list, while the complete top-4 PageRank list maintain a position in top-20 followers list, having 3 of them in top10. Kwak et al. observe in 2009 that although the two lists do not match exactly, users are ranked similarly by the number of followers and PageRank. However, from our 2015 study we conclude that the two rankings lists show significant differences. For example, Katy Perry has the most followers, but does not belong to the top-20 PageRank list, while 'CNN Breaking News', ranked 17th in followers, is ranked 3rd in PageRank. This fact could imply that Katy Perry's followers are mostly teenagers or average individuals with low PageRank, while 'CNN Breaking News' has many heavy-weight followers.

With these results we conclude that the number of followers does not provide us with strong insights regarding the topological influence of a user in the Twitter network.

6.4 Removed Users

Several studies have been performed on the characterization of the Twitter graph topology [58, 79] and users demographics [73]. Moreover, Liu et al. perform a study on the evolution of users behavior and highlight the rise of spammers and malicious behavior [66]. Thomas et al. analyze the behavior of suspended accounts and present insights regarding their OSN behavior [98].

In this section we present a study on the graph structure of users who were part of the graph in 2009 but do not belong in the Twittersphere anymore. In this study we consider all removed accounts and not only users who have been suspended from Twitter mechanism as in [98]. We divide these users in different groups, based on



(a) In-Degree (b) Out-Degree

Figure 6.10: In-degree and Out-degree of the 3 different categories of removed users.

Removed Reason	*Size*	*In-Degree*	*Out-Degree*
Banned from Twitter	1,042,060	61.77	263.68
Intentionally Left	4,365,923	28.12	28.88
Privacy Settings	179,800	21.37	23.48

Table 6.4: Sample size, Average In-degree and Out-degree values for the 3 different categories of removed users.

the reasoning behind their disappearance. We conclude to the following categories of users: (*i*) who intentionally removed their account, (*ii*) who updated their ego-network visibility settings to private, (*iii*) who have been banned from Twitter due to their OSN behavior (e.g. bots, spammers). In the rest of this section we describe the graph metrics of these groups and extract insights on the comparison between them.

6.4.1 Degree Distribution

Studying the degree distributions (Figure 6.10) enable us to extract insights regarding the position of a removed user in the network before its disappearance and correlate it with the reasoning behind the latter.

Table 6.4 presents the average numbers of the in-degree and out-degree for each one of the 3 categories. As we can derive, users who have been banned from Twitter maintain a much larger out-degree, about 9-times more than the other categories, while the value for the rest two categories differ by only 5.4 nodes on average. Furthermore, the latter two categories maintain a ratio between out-degree and in-degree of about 1, which is an indication that these were maintained mostly by

individuals. However, the case for the users who have been banned from Twitter monitoring services is completely different; despite the fact that these users maintain a larger value of in-degree than other categories, the out-degree/in-degree ratio has a value higher than 4.2.

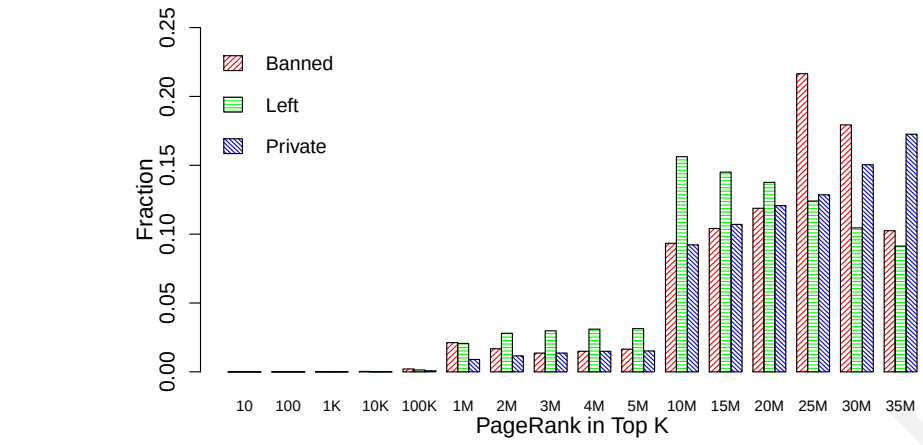
Discussion: From the results we can conclude that users who have been banned from Twitter have showed a degree distribution which has been observed on bots and/or spammers in past studies [20]. Regarding the rest two categories, we can see that they showed similar degree characteristics before their disappearance and can be related with an average non-active Twitter user. From these findings we can conclude that Twitter social graph eventually gets cleaned from non-active users, as they tend to disappear from the network either by intentionally removing their accounts or by maintaining strict privacy settings.

6.4.2 Connected Components

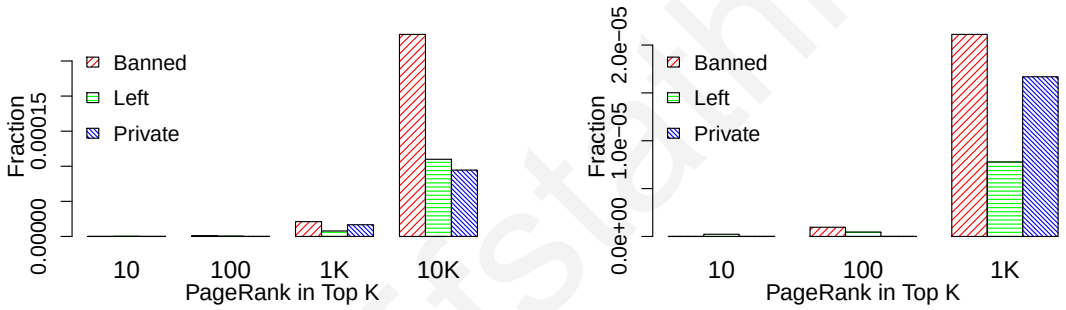
For the 3 categories we examine the Weakly and Strongly Connected Components that they participate in 2009. In the Weakly Connected Components (WCC) the direction of the connectivity is ignored, while in the Strongly is taken into consideration. Studying the connected components of the nodes who left twitter enable us to derive insights about their involvement in the social graph.

Regarding the users who have been banned from Twitter, 80% of them participate in the largest strongly connected component. Similarly, 81% of users who have update their privacy settings participate in the same Strongly Connected Component (SCC). However, the fraction of participation increases in the case of users who intentionally deactivate their profile, as 88% of them participates in the largest SCC. The percentages observed for the 3 categories of removed users are much higher than the corresponding values for the average twitter users presented in Section 6.2 and also observed by Myers et al. [79].

For the case of the Weakly Connected Components, we observe that in the 3 cases of removed users, all of them (100%) participate in the largest component. As presented in Section 6.2, more than 99.9% of the overall Twitter users participate in the largest WCC in the 2009 graph.



(a)



(b) (c)

Figure 6.11: PageRank in highest ranking lists for the 3 different categories of removed users.

6.4.3 PageRank

The degree distribution and connected components metrics give us an overview on the graph activity of a node in the network. In order to have a better overview on the influence of a node in the topology we use PageRank algorithm. We apply PageRank algorithm on the complete Twitter network of 2009 and extract insights regarding the different categories of removed users.

Figure 6.11 presents the fraction of removed users in the top rankings for each category. As we can see in Figure 6.11(a), the largest fraction of users who update their privacy settings (blue) appears in the lower ranking lists. Regarding the highest rankings, we observe that users who have been banned from Twitter hold the largest fraction until the top-2M list, were users who intentionally left take over.

Figures 6.11(c) and 6.11(b) present the results of the highest ranking lists. Surprisingly, we observe 60 users in total, who belong in the top-1K lists and are not part of the network today. The majority of these users have been banned from Twitter, while several others highly ranked users have changed their privacy settings. Furthermore, several users who intentionally left the network appeared in the highest rankings; 3 were in the top-100 lists, while one of them held a position in the top-10 higher ranked users of the network.

Chapter 7

Conclusions and Future Work

7.1 Distributed Large-Scale Data Collection in Online Social Networks

This study presents a framework for efficient data collection from Online Social Networks, enabled through crowd crawling of API data retrieval tokens. The proposed framework is based on the use of multiple OSN accounts, which are engaged in an efficient and smart distributed collection process, able to circumvent the imposed limitations without violating the terms of use. In all cases, the proposed solution proceeds to a pre-processing step where data are being anonymized, with respect to users' privacy and OSNs API terms-of-service. The evaluation of our proposed solution demonstrates its performance, in terms of dataset completeness and timeliness, for the case study of Twitter, one of the most popular platforms used in research. The presented framework enables the collection of more than 2.3M users in one day, retrieving also their Followers, Followees and Tweets. Furthermore, due to the intelligent use of resources, our framework triples the collection of the real-time stream of Twitter API.

7.2 Users Key Locations in Online Social Networks: Identification and Applications

We presented an effective methodology for the identification of a Twitter user Key locations. Our methodology uses geo-tagged Twitter data, and based on two main observations regarding user's real life habits, manages to identify the *Home* and

Work location of the users. Evaluation of our method, using data from several geographical regions, showed that it outperforms previous methods by more than 30%. Additionally, it can identify the user's Key locations at post-code granularity, that is in a radius smaller than 3Km. Comparison with socio-economic open data showed that our method can correctly identify the populated areas of the geographical region of interest.

To further evaluate our proposed methodology we illustrate how one can combine information from multiple social networks, namely LinkedIn and Twitter, in order to construct a dataset that includes both the user's Work location and her tweet activity. Using this dataset we evaluated our method for Work location identification. Our results show an accuracy close to 80% for identification of user location in a 10Km proximity. To the best of our knowledge this is the first attempt to construct a workplace ground truth dataset and also the first workplace identification method.

Furthermore, we briefly examine three different applications of location based social network analysis. Our results show the effect of locality in both the mobility patterns and the Ego network of Twitter user's. Also, we show that Key locations can be used to examine user's sentiment in more detail, taking into account small geographical areas. We believe that our initial results from this analysis will motivate further research in the area, aiding in better understanding of the behavior of Twitter user's both online and offline.

Our future work plans include the use of the identification derived from the methodology described in this work to perform a more thorough analysis of the applications illustrated in the previous section and derive insights for the users daily activities. Additionally, we aim in studying how the locations visited by the user affect her social network connections, and how the user transports derived by Twitter data can be used to support city planning procedures. Based on the insights that we extract from the location identification we aim in examining the influence of the culture in OSN activity and in the construction of the social graph. Furthermore, our future plans include the study of the influence of key locations in the graph constructed by the users that mention each other in their Tweets

7.3 Sentiment of Entrepreneurs in Twitter

This study has examined the sentiments (directed emotions) of entrepreneurs by comparison with those of non-entrepreneurs, as well as the sentiments of social and serial entrepreneurs. Building on the theory of entrepreneurial emotion, we developed hypotheses on the relations between the sentiments of these groups and tested the hypotheses using more than 29.5M messages sent by entrepreneurs and non-entrepreneurs on the social media website Twitter. We found that entrepreneurship can lead an entrepreneur to experience more positive general sentiment relative to non-entrepreneurs, while they experience less positive sentiment when discussing business matters.

We further showed that social entrepreneurship can lead the entrepreneur to experience more positive general sentiment relative to other entrepreneurs, while serial entrepreneurship can lead the entrepreneur to experience less positive general sentiment relative to other entrepreneurs. Our findings indicate that the differences in sentiment can be large – for example, the sentiment of entrepreneurs is typically 30 percent more positive than the sentiment of non-entrepreneurs with the same personal characteristics.

As a future work, we aim in studying the correlation between sentiment and weather conditions at the time that tweets have been published. We aim in investigating the influence that weather has on the sentiment of entrepreneurs, in comparison with the average non-entrepreneur user.

7.4 Online Social Network Evolution: Revisiting the Twitter Graph

In this study we revisit the Twitter network as it appeared in 2009 and re-collect the users full characteristics as of late-2015. In total we retrieve 34.66M users connected by 2.06B social connections. We perform a comprehensive study of the 2009 and 2015 social graph snapshots and present the results regarding various metrics in the topology of the social graph. In specific, we compare the two network snapshots and study the distributions of followers and followings, the relation between followers and tweets, reciprocity, degrees of separation, connected components and

differences in newly created and removed edges. Our results show a denser network with increased reciprocity but lower connectivity, as shown by the decrease in the networks largest strongly connected component. The average shortest path of the network also slightly decreases to 4.05 hops. We then examine the influential users of the network, as these can be defined by the number of followers and PageRank metrics. Our results show a significant change of these users between the years.

Having access to the entire 2009 Twittersphere, we identify users who do not belong in this directory anymore and investigate the reasoning behind their disappearance. We group removed users based on the reason they left the network and present a detailed comparison of the topological characteristics. We show that they have significant differences from the remaining set of users regarding their degree distributions, participation in Weakly and Strongly Connected Components, and their influential position in the social graph using their PageRank rankings. The results suggest that users who have been banned from Twitter showed different degree distributions than other categories, while the participation in WCC and SCC is much lower than the rest of the users. To the best of our knowledge this work is the first quantitative study on the entire Twittersphere, which compares the evolution of the network in such a large scale. We also introduce the study on removed users, where we group them in different fields and investigate their position in the social graph before their disappearance.

As future work we aim in studying users tweeting activity through the years and examine the correlation with the differences in graph topology. For 93% of the dataset we are able to access their complete public time-line, as these users posted less than 3,200 tweets [1]. For this large group we can study the different dynamics and influence of the tweeting activity, as it was before and after 2009. Our future plans include the construction and analysis of the re-tweets graphs for both 2009 and 2015 dataset snapshots. Furthermore, we aim in performing a study of the graph constructed by the users that mention each other in their Tweets, and compare the insights with the mutual relationships graphs.

[1] Twitter API enables the retrieval of at most 3,200 latest tweets per user.

Bibliography

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 285–294. ACM, 2012.

[2] S. Adali and J. Golbeck. Predicting personality with social behavior: a comparative study. *Social Network Analysis and Mining*, 4(1), 2014.

[3] A. Aspelund, T. Berg-Utby, and R. Skjevdal. Initial resources' influence on new venture survival: a longitudinal study of new technology-based firms. *Technovation*, 25(11):1337–1347, 2005.

[4] S. Asur and B. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499, Aug 2010.

[5] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 2011.

[6] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, pages 49–62, New York, NY, USA, 2009. ACM.

[7] M. Berlingerio, F. Calabrese, G. Di Lorenzo, X. Dong, Y. Gkoufas, and D. Mavroeidis. Safercity: A system for detecting and analyzing incidents from social media. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 1077–1080, Dec 2013.

[8] S. Bird. Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[9] A. Black, C. Mascaro, M. Gallagher, and S. P. Goggins. Twitter zombie: Architecture for capturing, socially transforming and analyzing the twittersphere. In *Proceedings of the 17th ACM International Conference on Supporting Group Work*, GROUP '12, pages 229–238, New York, NY, USA, 2012. ACM.

[10] M. Bošnjak, E. Oliveira, J. Martins, E. Mendes Rodrigues, and L. Sarmento. Twitterecho: A distributed focused crawler to support open research with twitter data. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 1233–1240, New York, NY, USA, 2012. ACM.

[11] A. Bozzon, H. Efstathiades, G.-J. Houben, and R.-J. Sips. A study of the online profile of enterprise users in professional social networks. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 487–492, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[12] C. Brown, A. Noulas, C. Mascolo, and V. Blondel. A place-focused model for social networks in cities. In *Social Computing (SocialCom), 2013 International Conference on*, pages 75–80, Sept 2013.

[13] J. Cao, H. Gao, L. Li, and B. Friedman. Enterprise social network analysis and modeling: A tale of two graphs. In *INFOCOM, 2013 Proceedings IEEE*, pages 2382–2390, April 2013.

[14] M. S. Cardon, J. Wincent, J. Singh, and M. Drnovsek. The nature and experience of entrepreneurial passion. *Academy of management Review*, 34(3):511–532, 2009.

[15] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the International Conference on WWW 2009*.

[16] C. M. Cheung, P.-Y. Chiu, and M. K. Lee. Online social networks: Why do students use facebook? *Computers in Human Behavior*, 27(4):1337–1343, 2011.

[17] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, New York, NY, USA, 2011. ACM.

[18] J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 124–135, New York, NY, USA, 2002. ACM.

[19] W.-H. Chong and E.-P. Lim. Tweet geolocation: Leveraging location, user and peer signals. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1279–1288, New York, NY, USA, 2017. ACM.

[20] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the ACSAC 2010*.

[21] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris. Assessing the potential of ride-sharing using mobile and social data: A tale of four cities. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 201–211, New York, NY, USA, 2014. ACM.

[22] P. A. Dacin, M. T. Dacin, and M. Matear. Social entrepreneurship: Why we don't need a new theory and how we move forward from here. *The academy of management perspectives*, 24(3):37–57, 2010.

[23] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 35–44, New York, NY, USA, 2010. ACM.

[24] M. D. Dikaiakos and D. Zeinalipour-Yazti. A distributed middleware infrastructure for personalized services. *Computer Communications*, 27(15):1464 – 1480, 2004.

[25] C. Ding, Y. Chen, and X. Fu. Crowd crawling: Towards collaborative data collection for large-scale online social networks. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 183–188, New York, NY, USA, 2013. ACM.

[26] C. Efstathiades, A. Belesiotis, D. Skoutas, and D. Pfoser. Similarity search on spatio-textual point sets. In *Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016, Bordeaux, France, March 15-16, 2016, Bordeaux, France, March 15-16, 2016.*, pages 329–340, 2016.

[27] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos. Identification of key locations based on online social network activity. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 218–225, New York, NY, USA, 2015. ACM.

[28] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos. Distributed large-scale data collection in online social networks. In *2016 IEEE International Conference on Collaboration and Internet Computing (CIC)*, pages 373–380, Nov 2016.

[29] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos. Users key locations in online social networks: identification and applications. *Social Network Analysis and Mining*, 6(1):1–17, 2016.

[30] H. Efstathiades, D. Antoniades, G. Pallis, M. D. Dikaiakos, Z. Szlávik, and R.-J. Sips. Online Social Network Evolution: Revisiting the Twitter Graph. In *2016 IEEE International Conference on Big Data*, IEEE BigData 2016, 2016. Best student paper award.

[31] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, Stroudsburg, PA, USA, 2010.

[32] D. Falcone, C. Mascolo, C. Comito, D. Talia, and J. Crowcroft. What is this place? inferring place categories through user patterns identification in geo-tagged tweets. In *Proceedings of International Conference on Mobile Computing, Applications and Services*, MobiCASE 2014.

[33] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *CoRR*, abs/1407.5225, 2014.

[34] L. Gabrielli, S. Rinzivillo, F. Ronzano, and D. Villatoro. From tweets to semantic trajectories: Mining anomalous urban mobility patterns. In J. Nin and D. Villatoro, editors, *Citizen in Sensor Networks*, Lecture Notes in Computer Science, pages 26–35. Springer International Publishing, 2014.

[35] P. Georgiev, A. Noulas, and C. Mascolo. The call of the crowd: Event participation in location-based social services. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014.

[36] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872–1892, October 2011.

[37] J. Gu, J. Tian, X. Wang, and H. Ling. Does negative news travel fast? exploring the effect of news sentiment on interactive spiral. In C. Stephanidis, editor, *HCI International 2017 – Posters' Extended Abstracts*, pages 435–442, Cham, 2017. Springer International Publishing.

[38] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Rec.*, 42(2):17–28, July 2013.

[39] W. T. Harbaugh, U. Mayr, and D. R. Burghart. Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, 316(5831):1622–1625, 2007.

[40] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[41] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

[42] E. Herder, P. Siehndel, and R. Kawase. Predicting user locations and trajectories. In *User Modeling, Adaptation, and Personalization*, pages 86–97. Springer, 2014.

[43] K. F. Hew. Students' and teachers' use of facebook. *Computers in Human Behavior*, 27(2):662–676, 2011.

[44] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1137–1146, New York, NY, USA, 2011. ACM.

[45] A. Iamnitchi, J. Blackburn, and N. Kourtellis. The social hourglass: An infrastructure for socially aware applications and services. *IEEE Internet Computing*, (3):13–23, 2012.

[46] M. Jones and D. Hardt. The oauth 2.0 authorization framework: Bearer token usage. Technical report, 2012.

[47] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding Human Mobility from Twitter. *ArXiv e-prints*, Dec. 2014.

[48] D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships, 2013.

[49] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the International Conference on WWW 2013*.

[50] S. Katragadda, M. Jin, and V. Raghavan. An unsupervised approach to identify location based on the content of user's tweet history. In *Active Media Technology*, volume 8610 of *Lecture Notes in Computer Science*, pages 311–323. Springer International Publishing, 2014.

[51] S. Katragadda, M. Jin, and V. Raghavan. An unsupervised approach to identify location based on the content of user's tweet history. In D. Ślęzak, G. Schaefer, S. Vuong, and Y.-S. Kim, editors, *Active Media Technology*, volume 8610 of *Lecture Notes in Computer Science*, pages 311–323. Springer International Publishing, 2014.

[52] J. Kaur and M. Bansal. *Hierarchical Sentiment Analysis Model for Automatic Review Classification for E-commerce Users*, pages 249–267. Springer International Publishing, Cham, 2017.

[53] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08, pages 19–24, New York, NY, USA, 2008. ACM.

[54] J. Kulshrestha, F. Kooti, A. Nikravesh, and P. K. Gummadi. Geographic dissection of the twitter network. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[55] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the ACM SIGKDD 2006*.

[56] S. Kumar, X. Hu, and H. Liu. A behavior analytics approach to identifying tweets from crisis regions. In *Proceedings of the ACM HT 2014*.

[57] S. Kumar, F. Morstatter, and H. Liu. *Twitter data analytics*. Springer, 2014.

[58] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the International Conference on WWW 2010*.

[59] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[60] A. Kyrola, G. Blelloch, and C. Guestrin. Graphchi: Large-scale graph computation on just a pc. In *USENIX Symposium on OSDI 2012*.

[61] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 915–924, New York, NY, USA, 2008. ACM.

[62] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1), Mar. 2007.

[63] G. Li, J. Hu, J. Feng, and K.-L. Tan. Effective location identification from microblogs. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 880–891, March 2014.

[64] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.

[65] Y.-R. Lin, B. Margolin, A. Keegan, A. Baronchelli, and D. Lazer. #bigbirds never die: Understanding social dynamics of emergent hashtag. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, 2013.

[66] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *ICWSM 2014*.

[67] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol.*, 5(3), July 2014.

[68] H. A. Maruf, N. Meshkat, M. E. Ali, and J. Mahmud. Human behaviour in different social medias: A case study of twitter and disqus. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 270–273, New York, NY, USA, 2015. ACM.

[69] M. L. Mauriello, C. Buntain, B. McNally, S. Bagalkotkar, S. Kushnir, and J. E. Froehlich. Smidgen: An approach for scalable, mixed-initiative dataset generation from online social networks. *SIGCHI*, 2018.

[70] F. T. McAndrew and H. S. Jeong. Who does what on facebook? age, sex, and relationship status as predictors of facebook use. *Computers in Human Behavior*, 28(6):2359–2365, 2012.

[71] A. Milani, N. Rajdeep, N. Mangal, R. K. Mudgal, and V. Franzoni. Sentiment extraction and classification for the analysis of users' interest in tweets. *International Journal of Web Information Systems*, 14(1):29–40, 2018.

[72] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[73] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist. Understanding the demographics of twitter users. In *ICWSM 2011*.

[74] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.

[75] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: A survey study of status message q&a behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1739–1748, New York, NY, USA, 2010. ACM.

[76] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.

[77] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*, 2013.

[78] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 913–924, New York, NY, USA, 2014. ACM.

[79] S. A. Myers, A. Sharma, P. Gupta, and J. Lin. Information network or social network?: The structure of the twitter follow graph. In *Proceedings of the International Conference on WWW 2014 Companion*.

[80] S. Narr, M. Hulfenhaus, and S. Albayrak. Language-independent twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML), LWA*, pages 12–14, 2012.

[81] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 8:1–8:7, New York, NY, USA, 2011. ACM.

[82] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[83] T. H. Nguyen, K. Shirai, and J. Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603 – 9611, 2015.

[84] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *Proc. of the 5th Int'l AAAI Conference on Weblogs and Social Media*, pages 570–573, 2011.

[85] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[86] S. C. Parker. Do serial entrepreneurs run successively better-performing businesses? *Journal of Business Venturing*, 28(5):652–666, 2013.

[87] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272, 2011.

[88] Z. Qingqing and Z. Chengzhi. Detecting dietary preference of social media users in china via sentiment analysis. *Proceedings of the Association for Information Science and Technology*, 54(1):523–527, 2017.

[89] D. Ruths, J. Pfeffer, et al. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

[90] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

[91] K. Ryoo and S. Moon. Inferring twitter user locations with 10 km accuracy. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 643–648, 2014.

[92] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 723–732, New York, NY, USA, 2012. ACM.

[93] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[94] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: Modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 355–358, New York, NY, USA, 2011. ACM.

[95] N. Spasojevic, Z. Li, A. Rao, and P. Bhattacharyya. When-to-post on social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 2127–2136, New York, NY, USA, 2015. ACM.

[96] K. Starbird and L. Palen. (how) will the revolution be retweeted?: Information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16, New York, NY, USA, 2012. ACM.

[97] K. Tao, C. Hauff, G. J. Houben, F. Abel, and G. Wachsmuth. Facilitating twitter data analytics: Platform, language and functionality. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 421–430, Oct 2014.

[98] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the ACM SIGCOMM Conference in IMC 2011*.

[99] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[100] A. H. Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10, July 2010.

[101] H. Wang, Y. Fu, Q. Wang, H. Yin, C. Du, and H. Xiong. A location-sentiment-aware recommender system for both home-town and out-of-town users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1135–1143, New York, NY, USA, 2017. ACM.

[102] J. Weng and B.-S. Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.

[103] J. S. White, J. N. Matthews, and J. L. Stacy. Coalmine: an experience in building a system for social media analytics, 2012.

[104] J. Wiklund, T. Baker, and D. Shepherd. The age-effect of financial indicators as buffers against the liability of newness. *Journal of business venturing*, 25(4):423–437, 2010.

[105] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer Systems*, EuroSys '09, pages 205–218, New York, NY, USA, 2009. ACM.

[106] H. Xie, X. Li, T. Wang, R. Y. Lau, T.-L. Wong, L. Chen, F. L. Wang, and Q. Li. Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. *Information Processing & Management*, 52(1):61–72, 2016.

[107] C. Yang, R. Harkreader, and G. Gu. Empirical evaluation and new design for fighting evolving twitter spammers. *Information Forensics and Security, IEEE Transactions on*, 8(8):1280–1293, Aug 2013.

[108] Y. Yu and X. Wang. World cup 2014 in the twitter world: A big data analysis of sentiments in u.s. sports fans' tweets. *Computers in Human Behavior*, 48:392 – 400, 2015.

[109] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: Discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, New York, NY, USA, 2013.

[110] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, MobiCom '14, pages 201–212, New York, NY, USA, 2014. ACM.