IMT School for Advanced Studies, Lucca

Lucca, Italy

Analysis of Polarized Communities in Online Social Networks

PhD Program in Computer, Decision, and Systems Science

XXIX Cycle

By

Mauro Coletto

2017

The dissertation of Mauro Coletto is approved.

Program Coordinator: Prof. Rocco De Nicola, IMT Lucca

Supervisor: Dott. Claudio Lucchese, ISTI-CNR Pisa

Co-supervisor: Prof. Rocco De Nicola, IMT Lucca

The dissertation of Mauro Coletto has been reviewed by:

Prof. Arjen P. de Vries, Radboud University

Prof. Francesco Bonchi, Pompeu Fabra University (Barcelona)

IMT School for Advanced Studies, Lucca

2017

Contents

List of Figures

List of Tables

Acknowledgements

I would like to deeply thank my supervisor, Dott. Claudio Lucchese, for the patient guidance, encouragement and advice he has provided throughout my time as a Ph.D. candidate. Moreover, among all the researchers who worked with me I want to explicitly mention Prof. Aristides Gionis, Dott. Luca Maria Aiello and Dott. Fabrizio Silvestri, for their time, their patience, their availability and their inspiring knowledge.

Vita

December 9, 1987	Born, Motta di Livenza (Treviso), Italy
2009	Bachelor Degree in Information Management Engineering Final mark: 110/110 cum laude University of Udine
2010	International Exchange Program University of Southern Denmark
2012	Master Degree in Information Management Engineering Final mark: 110/110 cum laude University of Udine
2013	Research Fellow University of Udine
2015	Academic Collaborator Yahoo Labs in London
2016	Academic Visitor Aalto University
2017	Research Fellow Ca' Foscari University of Venice

Publications

JOURNALS

1. Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., Quattrociocchi, W. (2015) Science vs conspiracy: collective narratives in the age of (mis) information. PLOS ONE.
2. Bellio, R., Coletto, M. (2015) Simple outlier labelling based on quantile regression, with application to the steelmaking process. Applied Stochastic Models in Business and Industry.

CONFERENCES

1. Coletto, M., Esuli, A., Lucchese, C., Muntean, C. I., Nardini, F. M., Perego, R, Renso, C. (2016) Sentiment-enhanced Multidimensional Analysis of Online Social Networks: Perception of the Mediterranean Refugees Crisis. In Workshop on Social Network Analysis Surveillance Technologies, co-located with IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM 2016, August 18-21, San Francisco, CA, US.
2. Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2016) Polarized User and Topic Tracking in Twitter. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 945-948). SIGIR 2016, July 17-21, Pisa, Italy.
3. Coletto, M., Aiello, L. M., Lucchese, C., Silvestri, F. (2016) On the Behaviour of Deviant Communities in Online Social Networks. In 10th International AAAI Conference on Web and Social Media. ICWSM 2016, May 17-20, Cologne, Germany.
4. Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2015). Electoral Predictions with Twitter: a Machine-Learning approach. In 6th Italian Information Retrieval Workshop. IIR 2015, May 25-26, Cagliari, Italy.
5. Coletto, M., Lucchese, C., Orlando, S., Perego, R., Chessa, A., Puliga, M. (2014) Electoral Predictions with Twitter: a Joint Machine Learning and Complex Network approach applied to an Italian case study. In International Conference on Computational Social Science. ICCSS 2015, June 8-11, Helsinki, Finland.

6. Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Quattrociocchi, W. (2014) Misinformation in the loop: the emergence of narratives in online social networks. In 13th Conference of the Italian chapter of AIS (Association for Information Systems). ITAIS 2014, November 21-22, Genova, Italy.
7. Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., Quattrociocchi, W. (2014) Sensing information-based communities in the age of misinformation. In European Conference on Complex Systems. ECCS 2014, September 22-26, Lucca, Italy.

BOOK CHAPTERS

1. Coletto, M., Lucchese, C. (August 2016) "Social-spatio-temporal Analysis of Topical and Polarised Communities in Online Social Networks" in Encyclopaedia of Social Network Analysis and Mining, edited by Alhajj R. and Rokne J.
2. Coletto, M. (May 2010) "Quality Management - ISO 9000" (Chapter 13) in La guida del Sole 24 Ore alla Qualità - GRUPPO 24 ORE, edited by Sartor M. and Mazzaro V.

TO BE PUBLISHED

1. Coletto, M., Garimella, K., Lucchese, C., Gionis, A. (2017). A motif-based approach for identifying controversy. Submitted to the 11th International Conference on Web and Social Media. ICWSM 2017, May 15-18, Montreal, Canada.
2. Coletto, M., Aiello, L. M., Lucchese, C., Silvestri, F. (2017). Adult Content Consumption in Online Social Networks. Submitted to Social Network Analysis and Mining, edited by Reda Alhajj (Springer).

MINOR CONTRIBUTIONS

1. Vol: Museo Facile. Progetto sperimentale di comunicazione e accessibilità culturale - Section: Esperienze a confronto - Stefania, C., Coletto, M. (2015) "La Sentiment Analysis per i musei 2.0: un approccio bottom-up per la conoscenza del pubblico." Studi e ricerche del Dipartimento di Lettere e Filosofia, University of Cassino.
2. Ritondale, M., Coletto, M., Caldarelli, G. (2015) Application of network analysis to the trade routes of antiquities passing through the pontine islands. CAA 2015, March 30-April 3, Siena, Italy.

Presentations

1. Coletto, M. "Pornography Consumption in Online Social Networks" at Ca' Foscari University of Venice (Venice, November 2016)
2. Coletto, M. "GYM seminar: Deviant Communities in Online Social Networks" at CNR ISTI (Pisa, October 2016)
3. Coletto, M. "Sentiment-enhanced Multidimensional Analysis of Online Social Networks: Perception of the Mediterranean Refugees Crisis" at ASONAM (San Francisco, August 2016)
4. Coletto, M. "Deviant Communities in Online Social Networks" at Aalto University (Helsinki, June 2016)
5. Coletto, M. "On the Behaviour of Deviant Communities in Online Social Networks" at ICWSM (Cologne, May 2016)
6. Coletto, M. "Introduction to Embeddings and their Application for a Label Propagation Task in a Social Network" at IMT (Lucca, December 2015)
7. Coletto, M. "Predictions of Political Elections with Twitter" at Yahoo Labs (London, June 2015)
8. Coletto, M. "Twitter for Political Predictions" at CNR ISTI, HPC Lab (Pisa, April 2015)
9. Coletto, M. "Information Extraction Models and Algorithms for Online Social Networks" at IMT (Lucca, November 2014)
10. Coletto, M. "Introduction to R" at CNR ISTI, HPC Lab (Pisa, November 2014)
11. Coletto, M. "Development of Dynamically Evolving and Self-adaptive Software" at IMT (Lucca, April 2014)
12. Big Data, a Challenge full of opportunities: Coletto, M. "New Business Intelligence technologies: some Business cases" at University of Udine (Udine, June 2013)

Abstract

Increasingly, people around the globe use Social Media (SM) - e.g. Facebook, Twitter, Tumblr, Flickr, Youtube - to publish multimedia content (posting), to share it (retweeting, reblogging or resharing), to reinforce it or not (liking, disliking, favoriting) and to discuss (through messages and comments) in order to be in contact with other users and to get informed about topics of interest. The world population is ≈ 7.4 billion people, among them ≈ 2.3 billion (31%) are active social media users (Global Web Index data, Jan 2016). In fact, these virtual contexts answer the human need of aggregation that nowadays is translated into digital bonds among peers all over the world, in addition to the traditional face-to-face relationships. Online Social Networks (OSNs), then, provide a space for user aggregation in groups, expressing opinions, accessing information, contributing to public debates, and participating in the formation of belief systems.

In this context, communities are built around different topics of interaction and polarized sub-groups often emerge by clustering different opinions and points of view. Such polarized sub-groups can be tracked and monitored over time in an automatic way and the analysis of their interactions is interesting to shed light on the human social behavior. Even though many studies have been devoted to understand different aspects of the social network structure and its function, such as, community structure (For10), information spreading (BRMA12), information seeking (KLPM10), link prediction (LNK07), etc., much less work is available on analyzing online discussions, user opinion and public debates.

In this doctoral dissertation we analyze the concept of *polariza-*

tion by looking at interactions among users in different Online Social Networks. *Polarization* is a social process whereby a social group is divided into sub-communities discussing different topics and having different opinions, goals and viewpoints, often conflicting and contrasting (Sun02; Ise86). We are interested in studying how and to what extend it is possible to extract information about *polarized communities* by automatically processing the data about interactions created in Online Social Networks. We present the state of the art and we propose a novel detecting method which allows to identify polarized groups, track them and monitor the topic evolution in the discussion among users of an OSN over time by classifing the keywords used in the messages exchanged. We show that it improves the state of the art and we describe case studies conducted particularly on Twitter (CLOP16; CGGL17).
The benefits in understanding user opinions are detailed in the first chapters. Moreover, we use the proposed methodology and alternatives in different application contexts: misinformation (BCD$^+$14a; BCD$^+$14b; BCD$^+$15), politics (CLOP16; CLOP15; CLO$^+$15), social behaviors (CALS16a; CALS16b), and migrations (CLM$^+$16).
A further application of opinion mining is the task of predicting user behavior. We discuss the limitations and the challenges related to this research area by looking at the context of political elections and by digging into a case study of electoral prediction. We believe that the analysis of *polarized communities* is OSNs can be used to predict collective social behavior, but major improvements in the field can be achieved by integrating several sources of information, such as traditional surveys, multiple Online Social Networks, demographic data, historical information, events, cyber-physical data.
Therefore, *polarization* is integrated in a framework of analysis with other dimensions (time, location) to explore social phenomena from a social media perspective. In particular, we

look at the possibility to understand European perception of the political refugees' crises by mining OSN data.
The concept of *polarization* is related to that of *controversy*. *Controversy* describes the interaction among two or more opponent polarized communities that discuss together, often with heated tones. For some highly controversial topics (e.g., politics, religion, ethics) even though users prefer to get informed though polarized content originated in the communities they belong to, they like to share their affiliations, believes, ideals, convictions with external users in order to persuade them in joining their belief system or supporting, criticizing an event, a group, a party or a specific person. Highly *polarization* does not always imply *controversy* and vice versa. We describe the recent literature about *controversy* detection and we propose a machine learning approach which takes into account features related to the social network and to conversational interaction patterns. The model is able to identify controversy in a conversation without any feature related to the content of the interaction. The features are deeply analyzed and the accuracy of the model is discussed.
We finally explore two opposite situations. The first is the formation of *echo chambers*, where a user gets informed and gives opinions in a self-contained group, whose members share a similar point of view. By analyzing communities in Facebook which consume news from scientific pages and from pages focused on conspiracy theories we confirm the hypothesis of cognitive closure of the users, weakening the idea of Social Media as a space for democratic collective intelligence. The second is the presence of *deviant communities*. Those are communities that emerge around what are usually referred to as *deviant behaviors* (CM15), conducts that are commonly considered inappropriate because they violate society's norms or moral standards. An example of deviant behavior is the pornography consumption, that is the focus of our

examination looking at content dissemination in Online Social Networks. *Deviant communities* are commonly considered segregated but we show that instead their content might spread far away in the Online Social Network. We analyze both situations with real case studies using Facebook, Flickr, and Tumblr data.

Our work is an initial study of *opinion polarization* on Online Social Networks with some in-depth analyses of specific topical user communities. It brings novel contributions in: i) characterizing communities through the perspective of user *polarization*; ii) proposing a novel method to classify *polarized users* and topic evolution over time; iii) understanding user behavior from a social media perspective; iv) integrating *polarization* with other variables (time, space) with the purpose of analyzing a social phenomenon; v) defining *controversy* and how to detect it regardless of the content; vi) describing how people aggregate and share information in various contexts. Different topical communities and several OSNs are described in the dissertation, providing a general overview of the investigation field and proposing contributions to the discussion and solutions. Our research questions are part of a broader research area which is called Computational Social Science. This new discipline - which is the frame of our thesis - is a new approach to social studies by mean of novel large-scale computational tools, merging Social Science with Computer Science and Machine Learning.

Chapter 1

Overview: human aggregation from the *physical* to the *online world*

1.1 Introduction

The understanding of interactions among individuals, both face-to-face and in a virtual context, is a crucial task in many disciplines: Sociology, Psychology, Anthropology, Computer Science, Linguistics, Marketing, and Business. People aggregate in groups and interact each other sharing and producing information, which can be retrieved and analyzed to understand the opinions and the behavior of both single individuals and of the whole community.

In virtual contexts, specifically, it is possible to automatize the data mining process, collecting insights from million of users. The massive datasets obtained (*big data*) are precious resources that allow to extract unprecedented knowledge regarding social behaviors. Therefore, in the last 10 years, the focus of social studies widely included the digital world, resulting in a *marriage* between traditional social science and computer science approaches. Novel network analysis techniques and large-scale computational approaches have been developed to analyze users and

communities in virtual platforms (WF94).
In the next paragraphs, we give an overview of the general context of this work. We initially discuss why people aggregate in groups and how an analogous need is registered in the digital world. We briefly describe the field of Social Network Analysis reporting some relevant results and how new approaches have been developed in order to include online contexts. In particular, we focus on the novel research area of Computational Social Science, which is the reference field of this thesis, and we underline its perspective in comparison with traditional approaches of Social Science and Computer Science.
Furthermore, we describe some Online Social Networks used in our studies, and we report recent results in the analysis of communities in this context. Most of the content of the current introductory chapter will be published in the Encyclopedia of Social Network Analysis and Mining (CL16). The aim of this chapter is to give a panoramic view on the importance of the study of communities in digital platforms.
In this doctoral dissertation, then, we focus on the analysis of digital communities in Online Social Networks (OSNs), studying *polarization* of users in terms of similar opinions given a topic of discussion. In the following chapters we propose novel detection algorithms and we apply computational methods to study *polarized communities* in their context, discussing potential application areas. Finally, we describe the contributions of our research to the field.

1.1.1 Need of aggregation

"Man is a social animal." (Aristotle)

The Greek philosopher Aristotle, more than 2300 years ago, claimed that the social nature of human beings pushes them in being organized in groups at different scales: family, tribe, and society.
Our lives depend constantly on other humans and our social attitude has always been a top wide area of investigation for anthropologists, sociologists and psychologists. Animals in general exhibit social behaviors, embedded for instance in the concepts of territory and dominance.

However, some social traits are exclusively proper of the human species, which is organized in a social network without analogous cases in the animal realm, mainly due to the typically human use of a complex language and rationality (BHR07).
Moreover, most of human desires are based on social life. In developed countries people have largely fulfilled psychological and safety needs – as they are classified by Maslow in (Mas43) – which are for instance the need of food, water, sleep and security. Beyond these needs, according to a pyramidal hierarchy, humans look for sense of belonging and love, esteem and finally for self-actualization. These needs all involve social interactions and hardly can be obtained in isolation. Humans, then, organize themselves into social structures by their nature.

1.1.2 The *physical world*

The analysis of social networks is an interdisciplinary academic field that emerged from Social Science, Statistics, and Graph Theory. Groups and social networks have been studied for decades.
One of the main findings, with several important implications, is the so-called *small-world effect*: people in social networks are all linked by short chains of acquaintances (WS98). The closeness of people in the society was initially quantified by Milgram in 6 hops distance according to his famous experiment (Mil67) set in US.
Social structures like groups and dyadic ties are scrutinized by scientists to study human behavior and social interactions. Moreover, by looking at interactions among people in social groups, researchers have pointed out the presence of *strong* and *weak ties*, which structure the network in tightly clustered communities, with different roles in information spreading (Gra73).
Social studies successfully defined several theoretical models able to explain the patterns observed in these structures (WF94). In fact, social network and community analysis is currently one of the approaches of contemporary Sociology, and is also employed in a number of other formal sciences.

In Social Science, a *group*, or a community, is defined as a set of two or more people who interact with one another, share similar traits, and collectively have a sense of belonging. The definition implies three main concepts which have been extensively debated: *interdependence*, *homophily* and *social identity*. Interdependence refers to mutual interactions among the community members; homophily instead is the tendency of individuals to associate and bond with similar others; while social identity is the human sense of *who I am* based on group memberships.
These characteristics shift the definition of community beyond the simplistic idea of a group as an aggregation of individuals and entail a degree of subjectivity which makes the task of identifying communities hard. Interdependence and homophily can be measured and they have been studied in a quantitatively way (ABS+12; BAX10). On the other hand, the concept of social identity, which has been extensively studied at first by H.Tajfel (Taj82), is hard to frame and has been object of investigation. The psychological idea of group membership as a matter of shared self-definition is predominant (Tur81), but the subjectivity related to this concept is hardly treatable within the computer science context, which bases its findings on cohesive interpersonal relationships, by looking at interaction patterns. The matching between sociological findings and a computational approach to quantify them is still a challenging area.

1.1.3 Computer science perspective

In Computer Science (CS) the term *community* is more frequently used than the term *group*, which is widely adopted in Social Science (SS)[1].
According to a Computer Science terminology the discovery of communities is related to the task of clustering the nodes of a graph used to represent the social network.
People are mapped into nodes of a graph, and edges are created according to their interactions. Borrowing tools from clustering and theoretical graph analysis, a number of techniques have thus been used to to detect communities in social networks (e.g., Girvan-Newman method (GN02),

[1] In this thesis we use indistinctly both terms.

Modularity-based method (NG04)). Therefore, community detection techniques have been largely employed in recent years to uncover the structure of complex social systems.
However, the "algorithmic communities" are totally defined on the basis of some graph properties, e.g., density, and discard the subjective concept of *identity* of community members. Such communities emerge from the data which encode interactions according to predefined quantifying rules, leading to the detection of groups of users which are not always aware of being members of them. Groups detected algorithmically (*detected groups*) do not correspond to user-generated groups (*declared groups*) as considered in Social Science. Attempts to evaluate this mismatch has been done in (Aie15).
In Computer Science, the possibility to learn from data is the basic concept of Machine Learning and Data Mining techniques. Compared with conventional computational models, Artificial Intelligence approach offers a wide range of decisive advantages for Social Science: theoretical knowledge does not have to be formulated *a priori*, but it is enclosed in the data and discovered through Machine Learning and it can be explained and justified *a posteriori* (Man96).
Consequently, Machine Learning is used in Social Science for both *theory-driven* and *data-driven* model building. In the case of the theory-driven approach, knowledge based modeling allows the translation of theory into evaluation in order to confirm the hypothesis or to investigate the logical properties of the theory. With the data-driven approach, instead, it is possible to discover novel theoretical mechanisms inductively (Man96). The data-driven approach can be described as a *paradigm shift* in the research practice, which is, following Kuhn's definition (Kuh62), a phenomenon in which an abrupt shift in values, goals, methods of the scientific community occurs (Cri14). Some successful stories (from spelling correction to face recognition, including question answering, machine translation, information retrieval) show how the data-driven approach - relying on machine learning technologies - is the winning one in many applications (Cri14). For the analysis of user interactions and aggregation in communities machine-learning models are crucial since they can integrate

hundreds of variables (temporal, spacial, content-based, network-based, emotion-based) and infer some meaningful correlations able to contribute in the discovery of the drivers of social processes.

1.1.4 Computational social science perspective

Recently a new discipline based on the quantitative understanding of the complex social systems (CR10) was born: Computational Social Science. Computational Social Science (CSS) is the bridge between Social Science (SS) and Computer Science (CS), based on the study of what is proper of social studies through computational techniques and approaches developed in the CS community.
Computational Social Science can benefit from the presence of huge volumes of data on society's everyday behavior due to the significant integration of technology into the people's life (CGB+12). Over the past 100 years, social science has generated a tremendous number of inconsistent and contradictory theories on individual and collective human behavior. Duncan J. Watts (Wat17) and many other notable researchers believe that the computational solution-oriented approaches of CSS can advance social science, reducing the emphasis on the advancement of theories over the solution of practical problems.

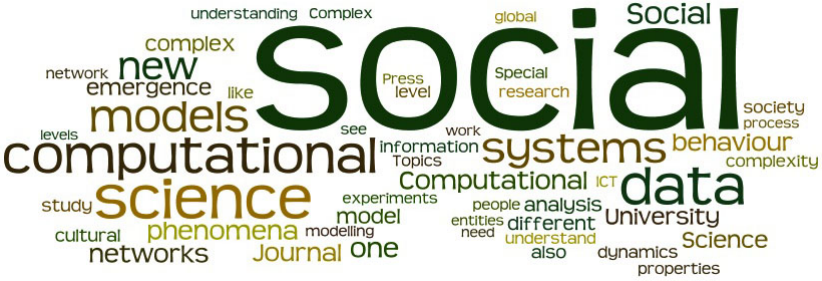


Figure 1: Word cloud of *Manifesto of Computational Social Science* - The most frequent terms contained in the document (CGB+12).

In Figure 1 the word cloud of the *Manifesto of Computational Social Science* (CGB$^+$12) is shown. The terms well describe the underlying concept which characterizes this novel discipline and define the frame of the current thesis work, which falls under this new field.

1.1.5 The *virtual world*

The birth of Online Social Networks (OSNs) in the late '90s and their increasing popularity in the early 2000s is an answer to the human need of belonging, even in the virtual world.
The success of Online Social Networks has been anticipated by the diffusion of virtual environments and the development of the web. In particular, virtual games have been precursors of Online Social Networks. In (MSL08) the authors describe the historical progression of the virtual world starting from arcade games, which started in 1972 with the Pong game by Atari Interactive. After that, the path towards OSNs was marked by the introduction of console systems (1986), followed by LAN Games, which created the concept of digital communities through Internet connectivity.
Game environments have progressively integrated additional social features with unstructured games and player generation of content (e.g., The Sims). Social networking sites are a further evolution in the development of open virtual worlds, which have properties that make them equivalent or at least comparable to the real world environments.

1.1.6 Online Social Networks

In an OSN an individual creates his own profile, publishes content and interacts with other users through discussions or actions (re-sharing content, liking, disliking). Users can also build friendships or subscriptions (following) links with other users.
The world population is ≈ 7.4 billion people, among them ≈ 3.4 billion (46%) are Internet users and ≈ 2.3 billion (31%) are active social media users (Global Web Index data, Jan 2016). These numbers suggest that there is a large interest in joining OSNs.

In our research and in the applications described in the following chapters of this dissertation we used data from various OSNs: Facebook, Twitter, Tumblr, and Flickr.

Facebook is the most famous Online Social Network with 1.71 billion monthly active users (Statista 2016). The Facebook website was launched on February 4, 2004, by Mark Zuckerberg, along with fellow Harvard College students and roommates. Data are not easily retrievable because of privacy limitations. In Chapter 5 we report a study based on data collected through this OSN.

Twitter is a micro-blogging platform. It was created in 2006 and it enables users to send and read short 140-character messages called "tweets". Twitter has 313 million monthly active users (Statista 2016) and it is the most used OSN in research since it was one of the first platforms that distributed the data through APIs. Tweets are public and the data collection is a quite easy task, even though many limitations have been introduced in the use of the APIs and in the scraping opportunities. In Chapters 2, 3, 4, and 6 we report experiments based on Twitter data.

Tumblr is an OSN founded in 2007 and owned by Yahoo! since 2013. Tumblr has 555 million monthly active users (Statista 2016) and it easy very popular among teenagers (in particular among females). Compared to other sociale networks Tumblr has less limitations in the content that can be published by the users and it has been widely used to spread adult content. In Chapter 7 we specifically look at the communities which produce this type of content and we study the dissemination.

Flickr is an image and video hosting service that was created in 2004 and acquired by Yahoo in 2005. It can be considered an OSN since it enables many social features (publishing, following, liking), but in comparison with the previously described OSNs it does not allow the user to internally share others' content. In Chapter 7 we specifically use data from this OSN to study adult-content-based communities, drawing a comparison with Tumblr.

In (OOL14) the authors show positive associations among the number of friends in OSNs, supportive interactions, affect, perceived social support, sense of community and life satisfaction. For this reason the time spent

Figure 2: Map of Facebook friendships - Friendship links in Facebook to visualize the massive amount of interactions in OSNs (2014).

by users in these networking platforms is significant: on average almost 2 hours per day according to the Global Web Index data.
Today, Online Social Networks represent a significant portion of the Web traffic and the pervasive use of these platforms together with the possibility to keep track of all actions have attracted scientists interested in investigating their properties. Such huge volume of information produced can be a gold-mine for researchers willing to investigate human behaviors in social environments, with no equivalent in the *physical world*.
The first studies on OSNs have regarded the topology and the structure of these large networks. From a topological point of view an OSN can be considered as a graph where nodes are users and edges are connections (friendships or following relations). Many works analyzed OSNs from a structural point of view, showing again a small world effect (BAA05), i.e., high clustering coefficient and short average path length (average degree of separation from 3 to 5) in different OSNs: Flickr, LiveJournal, Orkut, and YouTube (MMG$^+$07), Twitter (KLPM10), Facebook (UKBM11; WSPZ12; BBR$^+$12), Google+ (MCST$^+$12), studying the degree distribution which was found to be power law (UKBM11; WSPZ12) and

degree correlation, detecting the presence of a large strongly connected component (KNT10), finally investigating the evolution of graphs over time (WSPZ12).
Moreover, online social micro-blogging platforms and social networks have proven to be a rich source of information to track and monitor the behavior of users over time. Interactions in OSNs have been studied weighting the social graph through quantitative considerations on the strength of social ties. These graphs, called interaction graphs, differ from social graphs since they include quantifying mechanisms about the intensity of the connections, which dynamically changes. The interaction strength in a social network is a mix of amount of time spent together, intimacy, emotional intensity and reciprocal services (Gra73), but in most of the cases it is quantified in real OSN applications simply in terms of duration and frequency of contacts (e.g., in (WSPZ12)), even though there are theoretical studies, starting from (MC84), which try to translate qualities like intensity and intimacy into quantity values.
Ego networks are graphs where the central node is the studied user and all the other nodes connected to him represent his/her friends. Interaction graphs in OSNs have been studied showing both micro properties related to ego networks (looking for instance at close friends, inactive relationships, homophily, turnover of friendships) and macro properties related to the whole network (diameter, degree distribution, clustering coefficient) which are generally more stable (e.g., in Twitter(ACPD13), in Facebook (WSPZ12)).
In between ego networks and the whole social network there are clusters of users well connected: communities. These social structures have a salient role to study interests, opinions, influence, diffusion and many other social aspects which characterize users and their published content.

1.1.7 Communities in online social networks

Communities emerge around different topics of interaction and the analysis of the social aggregations in a virtual context is interesting to shed light on human behavior.

Homophily is a main driver that characterizes communities both in real and virtual contexts. Homophily induces similarity between members of communities: "*birds of a feather flock together*" (MSLC01). This is due to two co-founding principles: *i) selection mechanisms* and *ii) social contagion*. Selection mechanisms imply that preferences are connected to similar users' traits, while social contagion refers to how much linked people influence each other (Lee97).

Homophily has been widely studied in OSNs showing correlation between friendships and interests (ABS+12), or between profile information and communication patterns (LH08). Local proximity and age are another example of homophily factors in OSNs (KLNN+05).

On the other hand, it has been shown that diversity in the discussed topics or in the shared content favors the stability of a community as group members keep being stimulated by new input (LCFT04). Models of growth and longevity of groups in digital contexts have been also investigated (BHKL06).

The group size also affects the dynamics of interactions. The phenomenon has been deeply studied in real world social networks by Robin Dunbar (Dun92). He correlated the volume of the neocortex in primates with the amount of social stable relationships they have. He adapted the same theory to humans (Dun93), concluding that the amount of people with whom a person can maintain stable social relationships is about 150.

Similar results have been found in the Facebook friendship network, showing similarities between ego network structures in OSNs and in real life (ACPP12).

Similarly, Goncalves et at. (GPV11) performed comparable experiments in Twitter measuring the average interaction strength.

Two main processes can be identified in the development of communities in social networks: users create ties based on common interests, or based on personal social relationships.

The resulting kind of groups have been referred to as *common identity* vs. *common bond* (PML94). We also adopt the lexicon proposed by Martin-Borregon et al. (MBAG+14a), and refer to those groups as *topical* vs. *social*. Members of topical groups discuss a specific topic or a specific area of

interest and they do not usually have personal relationships between one another. Conversely, members of social groups tend to be reciprocal in the interactions with other members and discussions cover multiple topics.
One implication is that social groups are vulnerable to turnover, since personal relationships are present and they can influence user departure. Topical groups, on the other hand, are robust to departures and they are open to accept new members (Aie15). To discriminate groups Aiello suggested furthermore to look at specific variables (Aie15) to quantify the reciprocity of interactions and topical width of the discussions under a computational perspective. Typically larger reciprocity indicates a higher probability that the group is social, while a small topical width indicates topical groups. These variables integrate both social and content-based aspects.
The concept of *topic* in defining topical communities refers to a common interest among participants. From a computer science point of view, given a set of messages describing the interactions among users it is not easy to detect if the conversation regards many common interests, because often different topics share the same vocabulary. Indeed, in CS a topic is simply a multinomial distribution over words that represents a coherent concept in a text. To extract the most important topics from a piece of text (topic selection task) different techniques have been developed: one among the most popular methods is the unsupervised *latent Dirichlet allocation* (LDA) (BNJ03), which has been widely applied in different applications. Recently other advanced methods based on LDA have been proposed (MB08; BL07; WBS$^+$09). Even though CS provides sophisticated classification methods, the detection of topical groups is far from being an easy task. The understanding of natural language is a complex task and computational approaches show their limitations in understanding the variety of meanings and the underlying emotions provided by textual sources.
Furthermore, researchers have explored the relationship between diffusion of a topic and network structure (BBM13), focusing on the structural and dynamical properties of specific topical communities such as groups supporting political parties (CRF$^+$11a), or discussing various

conspiracy theories (BCD+15), rumors and hoaxes (RCM+11), deviant behaviors (CALS16a) or more ordinary topics like fashion or sports. These studies show that community structure is often topic dependent. In the final part of the thesis this concept is explored in detail by looking at the communities of pornography producers and their relationships with the rest of the social network in terms of interactions and content spreading. In practice, groups can be both *topical* and *social* and an additional level of aggregation is the user opinion. Users with similar belief systems tend to cluster together and this concept is highly explored in the rest of the thesis. In the next section we describe what is the contribution of our work to the field and the structure of the dissertation.

1.2 Contribution

In the previous section we described why people aggregate in groups and why the virtual context of OSNs is interesting for the understanding of user behavior.
We furthermore provided an introductory explanation about the importance of communities in OSN structure and how these groups have a more social or topical nature.
If we look specifically at the interactions, both in social and topical groups, users tend to start interacting about a common interest, an event or a specific content which initiates the discussion.
A fine-grained investigation of the communities reveals that there is a second level of clustering which is based on user opinion. When users discuss a specific topic, most of the times they share different points of view and they tend to associate according to their belief systems (BCD+15).
We call such addition level of clustering of an interactive community *polarization*. This concept is the main focus of the thesis. We investigate polarized sub-groups: aggregation of people who share opinions about a specific topic of interaction. In particular, if the topic is highly controversial, such sub-groups are strongly polarized. The relation between topic and polarization is dynamic: users generally start discussing about

a topic, and around that different opinions emerge.
Understanding opinion and polarization is a challenging task and it has recently received great attention in the Information Retrieval and in the Data Mining communities. In this work we dive into the problem of detecting, describing and analyzing polarized communities.
So far the content of interaction in OSNs has been studied mainly through the analysis of the *sentiment* of a specific portion of text. The sentiment is the general attitude of a speaker that is expressed in a message about the reference object of the discussion and usually it is labeled as positive, negative or neutral. Some sentiment analysis techniques provide a more wide scale of values that goes always from a very negative feeling to a very positive one.
The concept of polarization that we investigate is related with the sentiment of users, but it represents a wider class of elements, including the understanding of different points of view. Sentiment analysis, then, is a subset of polarization analysis.
To give an example let us consider a celebrity and users in an OSN discussing about him/her: there might be people supporting and liking him or her (i.e., expressing a positive sentiment), and people disliking and criticizing him or her (i.e., expressing a negative sentiment). These are polarized users but polarization is a wider concept and does not imply only positive or negative feelings but opinions in general. For instance, people might think that the celebrity is good in doing an activity (e.g., singing), but not in others (e.g., dressing) or they might believe or not in events happened to the celebrity, or again they might wish or not that something happens. All of these are points of view, opinions, desires and they can create polarization by fragmenting the community in people who share similar ideas.
Our goal is to design methods that are able to detect polarization, going beyond sentiment analysis techniques.
Detecting the sentiment from a text is a complex task, but is mostly related with natural language processing. In our context we would like to take into consideration polarization from a more general point of view by looking at the content, network features, temporal, spacial and

conversational patterns. Detecting, identifying and tracking polarization is helpful in understanding users opinions, preferences, and expectations. An additional contribution of our thesis in this field is to give a definition of polarization and to study how the concept can be adapted in specific application domains: misinformation (BCD$^+$14a; BCD$^+$14b; BCD$^+$15), politics (**?** CLOP15; CLO$^+$15), social phenomena (CLM$^+$16), and in particular among them deviant behaviors (CALS16a; CALS16b).
With specific case studies we investigate the concept of polarization among users, both in terms of communities focused on different topics and/or communities discussing a single topic with "clusterable" opinions, associating the concept of polarization with controversy. In the following chapters we focus on both concepts and we propose a methodology to detect both polarization and controversy, highlighting the differences between the two.

In particular, we want to explore the following research questions:

Q1: *How can we define polarization in OSNs? Can we automatically detect and track polarized users given the content of their interactions?*
In Chapter 2 we discuss in detail the concept of *polarization*. We present state-of-the-art methods that have been proposed to detect polarization and we describe a novel alternative method to identify polarized groups, track them and monitor the topic evolution in the discussion among users of an OSN over time (CLOP16).

- Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2016). Polarized User and Topic Tracking in Twitter. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 945-948). SIGIR 2016, July 17-21, Pisa, Italy.

Q2: *Can we predict user behavior trough the analysis of OSNs? Do OSNs give us insights to predict the outcome of a political election?*
The analysis of polarisation is useful to investigate the behavior of groups,

and it sometimes can be used to predict user activities, i.e., predicting vote intention among Twitter users (CLOP15) or understanding product preferences for marketing aims (LAH07).
In Chapter 3 we analyze one application domain where it is valuable to track polarization to detect user behavior: i.e., predicting voting behavior of users in Twitter.

- Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2015). Electoral Predictions with Twitter: a Machine-Learning approach. In 6th Italian Information Retrieval Workshop. IIR 2015, May 25-26, Cagliari, Italy.

- Coletto, M., Lucchese, C., Orlando, S., Perego, R., Chessa, A., Puliga, M. (2014) Electoral Predictions with Twitter: a Joint Machine Learning and Complex Network approach applied to an Italian case study. In International Conference on Computational Social Science. ICCSS 2015, June 8-11, Helsinki, Finland.

Q3: *Can we integrate polarization with other variables (i.e., time, space) to create an analytical framework that might be used to study social phenomena?*
Chapter 4 is dedicated to the integration of polarization with other variables in order to create a framework that can be used to analyze a social phenomenon (CLM$^+$16). We use the framework to analyze the recent issue of Mediterranean refugees and the perception of the phenomenon by European countries in Twitter.

- Coletto, M., Esuli, A., Lucchese, C., Muntean, C. I., Nardini, F. M., Perego, R, Renso, C. (2016). Sentiment-enhanced Multidimensional Analysis of Online Social Networks: Perception of the Mediterranean Refugees Crisis. In Workshop on Social Network Analysis Surveillance Technologies, co-located with IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM 2016, August 18-21, San Francisco, CA, US.

- Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2016). Polarized User and Topic Tracking in Twitter. In Proceedings of the 39th

International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 945-948). SIGIR 2016, July 17-21, Pisa, Italy.

Q4: *Do users interact with other users that do not share their belief system? How much isolated are communities in OSNs?*
In Chapter 5 we elaborate on *echo chambers* in Social Media and how people tend to share information, ideas, or beliefs inside an *enclosed* system, where different or competing views are censored, disallowed, or otherwise underrepresented. We analyze the case of communities of supporters of science and conspiracy theories in Facebook, how they are structured and to what extend users interact out of their own community (BCD+14a; BCD+14b; BCD+15).

- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., Quattrociocchi, W. (2015) Science vs conspiracy: collective narratives in the age of (mis) information. PLOS ONE
- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Quattrociocchi, W. (2014) Misinformation in the loop: the emergence of narratives in online social networks. In 13th Conference of the Italian chapter of AIS (Association for Information Systems). ITAIS 2014, November 21-22, Genova, Italy.
- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., Quattrociocchi, W. (2014) Sensing information-based communities in the age of misinformation. In European Conference on Complex Systems. ECCS 2014, September 22-26, Lucca, Italy.

Q5: *How do users of different polarized communities interact? What is controversy and how can me measure it? Can we automatically detect controversy without looking at the content delivered among users?*
People like to express their opinions in favor or against a particular idea, supporting or criticizing a particular political candidate or a party. In these

cases the polarized communities interact each other creating controversy. In Chapter 6 we explore the concept of *controversy* and, by presenting the state of the art in detecting methods, we discuss a novel approach which quantifies controversy without any information about the content of the conversation. The features used in the proposed machine-learning model take into consideration the social network, time-based actions, and - most importantly - conversational interaction patterns (CGGL17).

- Coletto, M., Garimella, K., Lucchese, C., Gionis, A. (2017). A motif-based approach for identifying controversy. Submitted to the 11th International Conference on Web and Social Media. ICWSM 2017, May 15-18, Montreal, Canada.

Q6: *How does content spread beyond niche or segregated communities?*
Even though in many contexts users are segregated in their *echo chambers* it might be that the produced content spreads through the weak ties or through the controversial interactions and goes beyond the producer communities. We explore this possibility in the last part of the thesis with a case study.
Chapter 7 is dedicated to the study of deviant communities, formation of topical communities centered on matters that are not commonly taken up by the general public because of the embarrassment, discomfort, or shock they may cause. These are polarized communities or at least topical communities usually considered very isolated from the rest of the social network. Since all these topics touch upon different societal taboos, the common-sense assumption is that they are embodied either in niches or in communities that might be quite numerous but whose activity runs separately from the mainstream social media life (CALS16a).
We show that for specific deviant communities, even though the producers are a small group, the content spreads far from the members who created it (CALS16a; CALS16b). Our analyses have been performed on Tumblr and on Flickr.

- Coletto, M., Aiello, L. M., Lucchese, C., Silvestri, F. (2016). On the Behaviour of Deviant Communities in Online Social Networks.

In 10th International AAAI Conference on Web and Social Media. ICWSM 2016, May 17-20, Cologne, Germany.

- Coletto, M., Aiello, L. M., Lucchese, C., Silvestri, F. (2017). Adult Content Consumption in Online Social Networks. Submitted to Social Network Analysis and Mining, edited by Reda Alhajj (Springer).

Other contributions developed during the Ph.D. and not included in the thesis are: a description of how sentiment analysis can be used in the context of Art Exhibitions and Museums (SC15); a statistical study on quantile regression (BC15); an introductory analysis of the benefits of network theory for maritime archaeology (RCC15).

Chapter 2

Polarization: detection and analysis

The results discussed in this chapter were published in (CLOP16).

- Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2016). Polarized User and Topic Tracking in Twitter. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 945-948). SIGIR 2016, July 17-21, Pisa, Italy.

2.1 Introduction

Digital traces of conversations in micro-blogging platforms and in Online Social Networks (OSNs) provide information about user opinion with a high degree of resolution. These information sources can be exploited to understand and monitor collective behaviors.
In this chapter, we study the concept of *polarization* in Social Media and, in particular, we propose a methodology to track *polarized communities* in a iterative way, also over time, to detect the dynamics of opinion polarization and to track the evolution of the topic of discussion by looking at keywords used in the messages exchanged by users.

The following is one of the definitions of polarization in Social Sciences.

> **Polarization**: *a social process whereby a social or political group is divided into opposing sub-groups having conflicting and contrasting opinions, goals and viewpoints (Sun02; Ise86).*

The number of opposing groups can be two (GMJCK13) or more, following a wider definition that we adopt. In the context of OSN we define *polarization classes* those topics that require the user to side exclusively with one position.

2.2 Problem formulation

In this chapter we aim at answering the following research question:
Q1: *How can we define polarization in OSNs? Can we automatically detect and track polarized users given the content of their interactions?*
In this chapter we discuss in detail the concept of *polarization*. We present state-of-the-art methods that have been proposed to detect polarization and we describe a novel alternative method to identify polarized groups, track them and monitor the topic evolution in the discussion among users of an OSN over time.

2.3 Contribution

In this chapter we describe a novel method to track users polarization and topics of the discussion in order not to forecast political events, which is a task highly exploited and undermined by difficulties (user bias, data dependency), but to monitor the *polarized* communities and their topics in the given data stream.
The proposed algorithm PTR (Polarization TRacker) provides an iterative classification of users and keywords: first, *polarized users* are identified, then *polarized keywords* are discovered by monitoring the activities of previously classified users. This method thus allows tracking users and topics over time.
While there exist several works about community detection and trending

topic tracking, we proposed a novel setting where the number of communities is known, very little information is provided to the algorithm (only a keyword per class), and those communities are competing each other. We report several experiments conducted on two Twitter datasets during political election time-frames. We measure the user classification accuracy on a *golden* set of users, and analyze the relevance of the extracted keywords for the ongoing political discussion.

2.4 Related work

In the recent years, the analysis of blogging platforms and streaming information sources (e.g., Twitter) has received great attention in the Information Retrieval and in the Data Mining communities. On-line social micro-blogging platforms have proven to be a rich source of information to track and monitor the behavior of users over time. Successful studies can be found in different contexts using such platforms for predictive tasks: from prediction of stock market (BMZ11) to movie sales (AH10), and pandemics detection (LDBC10). Other studies focus on the discovery of communities and trending discussion topics (LSM11; MK10).

We focus on the frequent scenario where users interact and produce contents according to a set of *polarization classes*. Political parties are typical examples of these classes. Users discuss about several parties, their opinion changes over time, but they can eventually vote only for one. Other examples include, for instance, brand analysis, products comparison, news discussions, and opinion mining in general.

According to such scenario, the polarization classes are known and some limited information may also be available, e.g., a set of relevant keywords. This limited knowledge allows to restrict the scope of the analysis, but several challenging tasks are left open.

The first is how to identify the users being polarized according to those classes (and the users not being polarized). The second challenging task is concerned with identifying the most relevant topics being discussed among such set of polarized users. The third is how to monitor the evolution of such user communities and their on-line discussions over

time. Those tasks are all very challenging as the available knowledge may be approximate or insufficient, and it may also become obsolete over time. Therefore, the classification into polarization classes should be able to self-update continuously by catching upcoming relevant users and discussion topics. In our experiments we used electoral data from Twitter. In this case, the polarization classes are political parties or candidates, and for each political formation at least one trivial hashtag is also known.
Several works analyzed the opportunities and limitations in using Twitter as a predictor of an election's outcome (DMBR13; TSSW10; CLOP15; GAMM11). In Chapter 3 we focus specifically on this task, proposing alternative predicting methods, but in this chapter our goal is different, as we do not draw any conclusion about the expected share of votes for the given parties or candidates. We use this specific typology of data, as they are a typical example of polarized users. We show that the proposed algorithm is able to identify candidate polarized users, by also analyzing the on going discussions among the respective communities. Our evaluation process is not related to electoral outcome, but we proposed an alternative method based on a control group, which can be used as a reference to measure the goodness of other methods that aim to classify the polarization of users according to a defined number of classes.
The present contribution is related to the Topic Detection and Tracking (TDT) subject (All12), which has been widely explored within the scope of news stream analysis (WJSS99). In particular we focus on content and user tracking for polarized users, which is connected with the concept of controversy in Social Media, which has been studied, mostly in political contexts, using data coming from different sources (blogs (AG05b), Twitter (CRF$^+$11b), Facebook(BCD$^+$15), news (MZDC14)). In Chapter 5 we explore more in detail the concept of controversy in Social Media and the different aspects of polarization.
Another related research area is trending topics analysis. A trend detection mechanism is proposed in (MK10), where bursty keywords are detected and then merged into groups on the basis of their co-occurrence. Finally, matrix factorization and entity extraction methods are employed to find a few representatives for each group. In (CDCS10) trendiness is defined

on the basis of the user authority and of the newly introduced keyword aging model which estimates the *energy* of a term over time, providing only a qualitative evaluation. The authors of (LSM11) use a language model based approach to attack the TDT problem specifically on Twitter data. They proposed to build a foreground and a background language model, respectively capturing recent past tweets, and different smoothing techniques are evaluated. Each topic is identified by a hashtag.
Similarly, in (YKSG14) a classification algorithm of tweets is proposed by mean of a hierarchy of topical categories. Training data are built by exploiting web page links in tweets, and only textual features are exploited.
Our approach is different in several regards from current literature and it is not aimed at classifying the topic of discussion of a given tweet or at detecting trending topics. We rather focus on the identification of polarized communities.
Some approaches based on networks analysis have been proposed to study polarization of users in social networks. For instance, in (CRF+11b) the authors studied the network of retweets and mentions to analyze the segregation of users using different clustering approaches. Recently (GDFMGM16) Garimella et al. proposed a graph-based method to identify controversy regarding topics analyzing properties of the partitioned graph of social interactions. An additional alternative approach to detect polarization is proposed in (LCN15), where the focus is on content of interaction and network data to infer user polarity.
In our case we want to study the polarization from a topical point of view, looking at the evolution of the discourse in terms of concepts discussed.

2.5 Data

We use two Twitter datasets related to political elections that recently took place in Italy.

Dataset IT13 Data about the primary election for largest social democratic political party in Italy (PD), which took place in December 2013

Table 1: Data statistics: full dataset

Dataset	**IT13**	**EU14**
tweets in original raw data	1.7 M.	2.3 M.
pre-electoral tweets $\mathcal{T}$	95,627	364,132
users with $\lvert\mathcal{H}_u\rvert > 0$	11,368 (65%)	28,340 (56%)

Table 2: Data statistics: golden dataset

Dataset IT13			**Dataset EU14**		
$\mathcal{C}$	Tweets	Users	$\mathcal{C}$	Tweets	Users
Renzi	330	109	PD	262	129
Cuperlo	4759	243	M5S	146	95
Civati	2925	700	FI	1263	199
			LN	480	226
			AET	757	328
total	8014	1052	*total*	2908	977

with 3 candidates: Mr. Renzi, Mr. Cuperlo, and Mr. Civati.

Dataset EU14 Data about the European Parliament election held in Italy in May 2014[1].
The data were collected through the Twitter API by querying a list of keywords related to the topic and the candidates, large enough to guarantee a good coverage of the elections. Both final datasets cover 9 days before the election day. We discard partial data and potentially irrelevant tweets, considering only tweets in Italian language. Table 2 reports some information about the two datasets.

2.6 Evaluation

We build an evaluation dataset by identifying those users whose opinion can be inferred with high confidence.

[1] The main national parties connected to different European political groups were: *Partito Democratico* (PD), *Movimento 5 Stelle* (M5S), *Forza Italia* (FI), *Lega Nord* (LN), *Tsipras* (AET). We ignore smaller parties and NCD-UDC for its limited presence in Twitter.

During elections, as for other events, very specific hashtags are used over Twitter to express a strong intention of vote or an explicit membership in a group.
We assume that users that frequently use one of such hashtags are strongly sided with one of the competing parties and they will not change idea in the short term. Such hashtags, named *golden hashtags*, are handpicked among the 500 most frequent in the data.
The golden hashtags are of the kind `#IVoteParty`. We identify one/two golden hashtags per class $c \in \mathcal{C}$ both in the EU14 (e.g., `#IVoteTsipras` for AET) and in the IT13 (e.g., `#prefeRenzi` for Renzi) dataset.
The set of reference users are identified by applying Algorithm 2 with the above *golden* hashtags as input. This guarantees that a user is safely considered as polarized to a party $c \in \mathcal{C}$ if her tweets contain only one of the golden hashtags associated with the various classes $c \in \mathcal{C}$. We denote with $\mathcal{Z} = \{z_1, z_2, \ldots\}$ this set of polarized users, and with $Z_c \subseteq \mathcal{Z}$ those supporting a specific formation c (Z_c is a partitioning of $\mathcal{Z}$).
The composition of resulting *golden dataset* is reported in Table 1. The golden dataset is thus a small fraction of the full dataset. A global analysis of the Twitter stream cannot be based on a few very polarized hashtags. Note that the relative popularity of the parties is not simply proportional to the number of votes received, but it depends on the efficacy of the hashtag promoted.
We remark that, for the sake of fairness, we remove the *golden* hashtags from the datasets before the application of any algorithm.
The set of users $\mathcal{Z}$ in the golden dataset, is used to evaluate the users classification accuracy of the proposed method. Given the users classification U_c provided by some given algorithm, precision, recall and F-Measure are restricted to the set $\mathcal{Z}$. Formally, for any given class $c \in \mathcal{C}$, precision and recall are defined as:

$$P_c(U_c) = \frac{|U_c \cap Z_c|}{|U_c \cap \mathcal{Z}|} \qquad R_c(U_c) = \frac{|U_c \cap Z_c|}{|Z_c|}$$

The F-measure F_c is the harmonic means of P_c and R_c. The macro F-measure average over the classes $c \in \mathcal{C}$ is denoted with F. In addition, as the proposed algorithm may not be able to classify all of the users in $\mathcal{Z}$,

we report also the user coverage γ and Γ on both the golden set and the overall dataset respectively:

$$\gamma(U = \cup_{c \in \mathcal{C}} U_c) = \frac{|U \cap \mathcal{Z}|}{|\mathcal{Z}|} \qquad \Gamma(U = \cup_{c \in \mathcal{C}} U_c) = \frac{|U|}{|\mathcal{U}|}$$

2.7 Method and algorithm

2.7.1 User and topic tracking

Let $\mathcal{T} = \{t_1, t_2, \ldots\}$ be the stream of tweets generated by the set of users $\mathcal{U} = \{u_1, u_2, \ldots\}$. We focus on the analysis of user behavior with respect to a set of *polarization classes* $\mathcal{C}$.
The goal of the proposed approach is thus to build a *partitional clustering* of the Twitter users, where each of the clusters is associated by construction with a single polarization class (or unassigned).
Our method can be seen as a *semi-supervised clustering* one, although, unlike classic methods, we do not provide any class representative around which the final clustering is induced. Indeed, the proposed method is only *loosely* supervised as the only knowledge available is the number of classes, and a short class description (a keyword).
An important issue is the evaluation of our algorithm. To this end, we exploit a *golden set* of polarized users, each unequivocally associated with a class $c \in \mathcal{C}$. Note that such knowledge is not exploited to train a classifier, but only for evaluation purpose.

2.7.2 The PTR algorithm

The Polarization TRacker (PTR) algorithm requires some initial *seed topics* that identify the classes of interests.
We propose to identify them with a single textual keyword for each class $c \in \mathcal{C}$. Although each keyword identifies a topic, e.g., a political party, it is not sufficient to correctly classify users, as all these seed topics are likely to be mentioned in many users' tweets, e.g., to contrast the achievements of a given party with the deficiencies of the others. We limit our keyword

Table 3: Notation

$\mathcal{T}$	the stream of tweets	Z	polarized users in golden set
$\mathcal{U}$	users posting $\mathcal{T}$	U_c, Z_c, H_c	set of elements classified as c
$\mathcal{H}$	hashtags mentioned in $\mathcal{T}$	$\mathcal{T}_u$, $\mathcal{H}_u$	tweets and hashtags by user u
t	a generic tweet in $\mathcal{T}$	U_c^*, H_c^*	candidate elements for class c
h	a generic hashtag in $\mathcal{H}$	$\mathcal{C}$	polarization classes c

selection to Twitter *hashtags*.

Therefore, the single textual keyword we initially choose for each class c is a single hashtag appearing in the user tweets, and around them we start identifying the user clusters.

The final goal is to extract the best discriminating hashtags that are able to identify the actual clusters of polarized users, who belong with high probability to one of the classes $c \in \mathcal{C}$.

We denote the representative hashtags, one for each $c \in \mathcal{C}$, called *seed hashtags*, by $H_c^{\tau=0}$, where τ is the algorithm's iteration number. Note that

Algorithm 1 PTR Algorithm

Require: The set of users $\mathcal{U}$ and their tweets $\mathcal{T}$ with hashtags $\mathcal{H}$, a single hashtag H_c^0 for each class $c \in \mathcal{C}$

Ensure: Classification of users U_c and hashtags H_c

procedure $PTR(\ \{H_c^0\}_{c\in\mathcal{C}}\ )$
  $\tau \leftarrow 0$
  for $c \in \mathcal{C}$ **do**
    $U_c^\tau \leftarrow \emptyset$
  end for
  repeat
    ▷ Classify users on the basis of the hashtags used
    $\left\{U_c^{\tau+1}\right\}_{c\in\mathcal{C}} \leftarrow \textsc{UserClass}\left(\{H_c^\tau\}_{c\in\mathcal{C}}, \{U_c^\tau\}_{c\in\mathcal{C}}\right)$
    ▷ Find better hashtags on the basis of $U_c^{\tau+1}$
    $\{H_c^{\tau+1}\}_{c\in\mathcal{C}} \leftarrow \textsc{HashtagsClass}(\{U_c^{\tau+1}\}_{c\in\mathcal{C}})$
    $\tau \leftarrow \tau + 1$
  until *convergence*
  return $\{U_c^\tau\}_{c\in\mathcal{C}}, \{H_c^\tau\}_{c\in\mathcal{C}}$
end procedure

Algorithm 2 User Classification Algorithm

Require: The set of polarized hashtags H_c and the previously found set of polarized users U_c^* for each class $c \in \mathcal{C}$
Ensure: New set of polarized users $\{U_c\}_{c \in \mathcal{C}}$

procedure USERSCLASS($\{H_c\}_{c \in \mathcal{C}}, \{U_c^*\}_{c \in \mathcal{C}}$)
  for $u \in \mathcal{U}$, $c \in \mathcal{C}$ **do** ▷ Find polarized tweets
    $T_{u,c} = \{t \in \mathcal{T}_u \mid \mathcal{H}_t \cap H_c \neq \emptyset \wedge \mathcal{H}_t \cap H_{c' \neq c} = \emptyset\}$
  end for
  for $c \in \mathcal{C}$ **do**
    $U_c \leftarrow \emptyset$
  end for
  for $u \in \mathcal{U}$ **do** ▷ Check user's polarization
    if $\exists c \in \mathcal{C} \mid \forall c' \in \mathcal{C}, c' \neq c \quad |T_{u,c}| > \alpha \cdot |T_{u,c'}|$ **then**
      $U_c \leftarrow U_c \cup u$
    else if $\exists c \in \mathcal{C} \mid u \in U_c^*$ **then**
      $U_c \leftarrow U_c \cup u$
    end if
  end for
  return $\{U_c\}_{c \in \mathcal{C}}$
end procedure

each initial set $H_c^{\tau=0}$, one for each c, is not necessarily composed of a discriminating hashtag. This set $H_c^{\tau=0}$ is then used to classify polarized users on the basis of their use of the seed hashtags. We denote by $U_c^{\tau+1}$ the clusters of users in $\mathcal{U}$ that are identified as belonging to class c, according to their tweets and to the given hashtags H_c^{τ}. Similarly, the new hashtags $H_c^{\tau+1}$ are generated by finding those that best discriminate the users in $U_c^{\tau+1}$. This refinement process is iterated for all $c \in \mathcal{C}$: from hashtags $\{H_c^{\tau}\}_{c \in \mathcal{C}}$ to users $\{U_c^{\tau+1}\}_{c \in \mathcal{C}}$, and finally to hashtags $\{H_c^{\tau+1}\}_{c \in \mathcal{C}}$. The algorithm terminates when H_c^{τ} converges. Algorithm 1 iterates two classification steps: classification of the users (USERCLASS) and classification of the hashtags (HASHTAGSCLASS).

Algorithm 2 illustrates the former step of the iterative process[2]. The goal of this step is to identify polarized users on the basis of the given hashtags. First, we identify polarized tweets, which mention hashtags in

[2]Note that we omitted the superscript τ for the sake of simplifying the notation.

Algorithm 3 Hashtag Classification Algorithm

Require: The set of polarized users U_c for each class $c \in \mathcal{C}$
Ensure: Polarized hashtags H_c

1: **procedure** HASHTAGSCLASS($\{U_c\}_{c \in \mathcal{C}}$)
2: **for** $c \in \mathcal{C}$ **do**
3: $H_c \leftarrow \emptyset$
4: $H_c^* \leftarrow \bigcup_{u \in U_c} \mathcal{H}_u$
5: **end for**
6: **for** $h \in \bigcup_{c \in \mathcal{C}} H_c^*$ **do**
7: **if** $\exists c \mid \forall c' \neq c \quad S_c(h) > \beta \cdot S_{c'}(h)$ **then**
8: $H_c \leftarrow H_c \cup h$
9: **end if**
10: **end for**
11: **return** $\{H_c\}_{c \in \mathcal{C}}$
12: **end procedure**

H_c. We consider the classification of each single tweet t by considering all the mentioned hashtags $\mathcal{H}_t$, as we believe each tweet is a very relevant expression of a user's thought on a specific topic.
Since we are interested in polarized users, with the goal of achieving high precision we discard all the tweets which contain hashtags belonging to more than one set $\{H_c\}_{c \in \mathcal{C}}$. For each user $u \in \mathcal{U}$ and for each class $c \in \mathcal{C}$ we denote the set of polarized tweets by $T_{u,c}$.
We thus measure the user polarization: if for some classes c, the number of tweets in $T_{u,c}$ is significantly larger than for any other class (parameter α), then the user is labeled with the class c and added to the set of polarized users U_c (see line 9). Note that the user classification is intended to be an update of the classification conducted during the previous step.
The goal the second step is to process all the hashtags adopted by classified users U_c in order to discover a new set of discriminating hashtags H_c, as illustrated in Alg. 3. In order to detect $\{H_c\}_{c \in \mathcal{C}}$, we take into considerations all the hashtags $\mathcal{H}_u$ used by any user $u \in U_c$, and not only those occurring in the polarized tweets $T_{u,c}$ (line 4). This allows to extend our analysis to the full set of topics discussed by the users, even if they were not captured in the early iterations of the algorithm. First, for each

$c \in C$ we retrieve the set of hashtags used by the users in U_c, considering all their tweets, denoted by $\mathcal{T}_c$, independent of the classification of the single tweets in the previous iteration. In our experiments we consider the top frequent 500 hashtags in $\mathcal{T}_c$.
Given the resulting set of candidate hashtags for each $c \in \mathcal{C}$, namely H_c^*, we extract from them the new hashtags that highly discriminate each class c, and these are eventually added to the new set H_c (line 8). Specifically, the discriminating hashtags are those highly used by the current set of users U_c, and partially used by any other user in $U_{c'}$, $c' \neq c$.
We define a function $S_c(h)$ to measure the goodness of hashtag h for each community of polarized users U_c. Let $\mathcal{T}_h$ be the set of tweets in $\mathcal{T}$ mentioning hashtag h, independent of the users who posted these tweets. Moreover, let $\mathcal{T}_{H_c^*}$ be the set of tweets in $\mathcal{T}$ containing at least one hashtag in the set H_c^*. We score the goodness of a hashtag for a polarization class as follows:

$$S_c(h) = \frac{|\mathcal{T}_h \cap \mathcal{T}_{H_c^*}|}{|\mathcal{T}_{H_c^*}|} \cdot \prod_{c' \in \mathcal{C}, c' \neq c} \left(1 - \frac{|\mathcal{T}_h \cap \mathcal{T}_{H_{c'}^*}|}{|\mathcal{T}_{H_{c'}^*}|}\right)$$

where we consider the naive hypothesis of independent occurrence of the hashtags in the various sets. In practice, $S_c(h)$ is the probability of seeing h only in H_c^*, whereas h is not present in all the other sets of hashtags $H_{c' \neq c}^*$.
Given a hashtag h, the score $S_c(h)$ is used to rank the various classes, thus assigning h to class with the highest score. Since we aim at promoting highly discriminating hashtags, not only we assign the hashtag h having the highest $S_c(h)$ to the new set H_c, but only if $S_c(h) > \beta \cdot S_{c'}(h)$, $\forall c' \neq c$, where $\beta \geq 1$. Note that if a tie exists between the to 2-top scores classes, the hashtag h is not assigned to any H_c, since it is considered not discriminating enough.

2.7.3 Baseline

As a baseline we use the k-means clustering algorithm. Each user u is represented by a vector of 500 features, corresponding to the 500

Table 4: Comparison with the baseline: k-means

Dataset IT13				Dataset EU14			
$\mathcal{C}$	P_c	R_c	F_c	$\mathcal{C}$	P_c	R_c	F_c
Renzi	0.144	0.257	0.185	PD	0.536	0.457	0.493
Cuperlo	0.252	0.543	0.344	M5S	0.359	0.895	0.512
Civati	0.766	0.366	0.495	FI	0.495	0.734	0.591
				LN	0.995	0.916	0.954
				AET	1.000	0.387	0.558
avg.	0.387	0.389	0.341	*avg.*	0.677	0.678	0.622
$\gamma = 1.0$		$\Gamma = 0.653$		$\gamma = 1.0$		$\Gamma = 0.557$	

most frequent hashtags in the dataset. The user feature vector stores the frequency of a hashtag in the stream of tweets $\mathcal{T}_u$ published by the user. We discard users who do not use any hashtag in their tweets.
We normalize the feature vectors for each user to unit L^2 norm. We impose the number of the clusters k equal to the number of classes $|\mathcal{C}|$ and, to simulate the same starting condition of our method, we built the initial centroids so as to encode the *seed* hashtags. The centroid for a class c is thus a vector with a single 1 in the position of the seed hashtag, and 0 otherwise. The result of the k-means baseline is thus a clustering of users based on the *seed* hashtags provided.
Table 4 reports the results of the k-means baseline. F-measure values are low for the IT13 dataset. For instance, k-means provides low accuracy and recall for the first class. This is mainly due to the fact that the hashtags corresponding to popular parties or candidates are very often used by different users, regardless of their orientation. In other cases (e.g., LN and AET), the hashtags are used mostly within the respective communities.

2.7.4 Results

In the following, we analyze in detail the iteration-by-iteration behavior of the proposed PTR algorithm. We test our algorithm by setting $\alpha = 2$ and $\beta = 1$, after a tuning step. During the first iteration, PTR is fed with the *seed* hashtags. Algorithm 2 uses those hashtags to find a subset of

Table 5: PTR Iteration-2 performance

Dataset IT13				Dataset EU14			
$\mathcal{C}$	P_c	R_c	F_c	$\mathcal{C}$	P_c	R_c	F_c
Renzi	0.350	0.752	0.478	PD	0.733	0.488	0.586
Cuperlo	0.869	0.300	0.446	M5S	0.325	0.842	0.469
Civati	0.916	0.747	0.823	FI	0.955	0.533	0.684
				LN	0.981	0.938	0.959
				AET	0.974	0.451	0.617
avg.	0.712	0.600	0.582	*avg.*	0.794	0.650	0.663
$\gamma = 0.845$		$\Gamma = 0.532$		$\gamma = 0.830$		$\Gamma = 0.367$	

polarized users in $\mathcal{U}$.

This step is similar to other works, where mentions of a party or candidate are used to estimate their popularity or to classify users (CLOP15; TSSW10). Unlike other approaches, PTR aims at discovering a subset of polarized users, thus requiring that a user mentions a party at least twice any other. The results of such user classification are evaluated over the *golden dataset*, as reported in the first line of Table 2.7.4. Regarding average precision, PTR is already significantly superior to the k-means baseline for IT13 dataset. This is already surprising, as the seed hashtags are very generic. On the other hand, the k-means baseline might be negatively affected by the sparsity of the data. The results are different on the two datasets in terms of average recall. PTR has similar performance to k-means on the IT13 dataset, while the recall is significantly lower on the EU14 dataset. This is confirmed by the coverage values γ and Γ.

In comparison with the baseline, the performance of PTR in terms of macro F-measure is satisfactory on the IT13 dataset, but not on the EU14 dataset yet. The output of the first iteration is a new set of hashtags which is exploited in the next iteration. By looking at the best scoring hashtag, we can already observe an interesting behavior of the algorithm for some $c \in \mathcal{C}$. In dataset EU14, the best tags for FI and LN are the leaders of the respective parties, detecting that the original *seed* hashtags are not discriminating in this case.

In Table 5 we report in detail the results after the second iteration of PTR.

Table 6: PTR iteration by iteration performance

	Dataset IT13			Dataset EU14		
Iter	F	γ	Γ	F	γ	Γ
1	0.358	0.490	0.218	0.514	0.670	0.163
2	0.582	0.845	0.522	0.663	0.830	0.367
3	0.588	0.853	0.532	0.662	0.831	0.386
4	0.588	0.853	0.534	0.661	0.834	0.390

Table 7: $TPTR$ day by day performance

	Dataset IT13			Dataset EU14		
Day	F	γ	Γ	F	γ	Γ
1	0.177	0.199	0.045	0.155	0.164	0.025
2	0.225	0.348	0.114	0.464	0.465	0.079
3	0.304	0.457	0.166	0.529	0.570	0.116
4	0.333	0.563	0.234	0.585	0.671	0.180
5	0.368	0.606	0.261	0.588	0.726	0.235
6	0.397	0.671	0.315	0.574	0.762	0.269
7	0.387	0.721	0.363	0.596	0.794	0.302
8	0.387	0.765	0.408	0.637	0.846	0.334
9	0.391	0.811	0.461	0.635	0.876	0.349

The first interesting result is that the average recall is significantly higher on both datasets. This is due to the new hashtags discovered in addition to the *seed* ones during the previous iteration, which, in turn, lead to the identification of a larger set of users: the coverage γ is now beyond 80% of the *golden* set, and Γ has doubled in this iteration.
Also the average precision is higher w.r.t. the previous iteration scoring more than 0.7. This is both because of the increased number of classified users, and of the updated user classification. As a result, the F-measure has an overall improvement w.r.t. the k-means baseline of +71% and +7% on datasets IT13 and EU14 respectively.
As shown in Table 2.7.4 PTR becomes stable very early. The largest improvement is achieved with the second iterations. This means that the most relevant hashtags are discovered early, and only slight changes occur

afterwards. The subsequent iterations marginally increase the number of classified users. Note that the algorithm is classifying the polarized users found in the whole set $\mathcal{U}$. PTR found about 6.7 and 27 thousands polarized users on the dataset IT13 and EU14 respectively. We conclude that in most cases, two iterations of the algorithm provide sufficient classification quality.
We do not report an exhaustive qualitative analysis of the outcome, but we observe that the procedure is able to extract relevant keywords: namely prominent politicians, the party itself and political mottoes characterizing each c in the political scene.
We finally propose a variant of PTR, that is $TPTR$ (temporal PTR), to perform the tracking of topics and users in time. In our case we consider the evolution day by day. The procedure follows Algorithm 2 and Algorithm 3 with the difference that at iteration τ only the tweets $\mathcal{T}_u$ written in the τ-th day are considered. We perform $TPTR$ on IT13 and on EU14 datasets. In Table 2.7.4 the evaluation of the temporal iterative procedure is shown. The macro F-measure is increasing day by day both for the effect of a better classification and for the presence of new users. Note that we evaluate the time iterative method day by day on the entire *golden set* of users. F-measure values are low because not all users in the *golden set* were active every day.

2.8 Conclusion

We proposed a novel algorithm for the simultaneous tracking of *polarized communities* and *discriminating topics* in OSNs. Specifically, it iteratively detects polarized users, and from their contents the discussed discriminating topics.
We also introduced a temporal variant, where the information extracted during one day of analysis is exploited for the next day. Indeed, the classification of users makes the algorithm more robust in terms of concept drifts, as new trends may be detected as early as they pop up. At the same time, the identification of discriminating topics helps in detecting users moving from one class to another.

The algorithm was tested on two Twitter data samples. We evaluated the quality of user classification on a *golden set* of users, showing significant improvements over the baseline. The proposed methodology is general and it can be applied to different scenarios.
We believe that this methodology based on *polarization* may also impact on broad area of social network analysis, e.g., by complementing the proposed classification with community detection and information diffusion over time.

Chapter 3

Prediction of user behavior in political elections

The results discussed in this chapter were published in (CLOP15; CLO$^+$15; CLOP16).

- Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2015). Electoral Predictions with Twitter: a Machine-Learning approach. In 6th Italian Information Retrieval Workshop. IIR 2015, May 25-26, Cagliari, Italy.
- Coletto, M., Lucchese, C., Orlando, S., Perego, R., Chessa, A., Puliga, M. (2014) Electoral Predictions with Twitter: a Joint Machine Learning and Complex Network approach applied to an Italian case study. In International Conference on Computational Social Science. ICCSS 2015, June 8-11, Helsinki, Finland.

3.1 Introduction

Chapter 2 discusses the concept of *polarization* and how we can automatically track opinions and polarized users in an Online Social Network (OSN) over time. This task is important because it can be used in different applications to predict user behavior.

The use of Social Media as a tool to predict the outcomes of social phenomena is a recurrent task in the recent social network analysis literature. In particular, one of the OSNs and micro-blogging platforms most used in research is Twitter, since it allows accessing the data through general APIs and special research agreements. Successful computational social studies can be found in different contexts using Twitter for predictive tasks: from prediction of stock market (BMZ11) to movie sales (AH10), and pandemics detection (LDBC10).
Computational Social Science (CSS) is becoming a leading research area in understanding communication patterns and social behaviors, in tracking tastes and, therefore, in predicting opinions (LPA$^+$09).
One application area of OSN opinion mining is the prediction of political electoral outcome. In this chapter we focus on the possibility to understand political orientations of users in Twitter in order to predict his/her vote intentions.

3.2 Problem formulation

In this chapter we aim at answering the following research question:
Q2: *Can we predict user behavior trough the analysis of OSNs? Do OSNs give us insights to predict the outcome of a political election?*
The analysis of polarisation is useful to investigate the behavior of groups, and it sometimes can be used to predict user activities, i.e., predicting vote intention among Twitter users (CLOP15) or understanding product preferences for marketing aims (LAH07).
In this chapter we analyze one application domain where it is valuable to track polarization to detect user behavior: i.e., predicting voting behavior of users in Twitter.

3.3 Contribution

Several studies have shown how to approximately predict public opinion, such as in political elections, by analyzing user activities in blogging platforms and on-line social networks. The task is challenging for several

reasons. Sample bias and automatic understanding of textual content are two of several non trivial issues.
In this chapter we study how Twitter can provide some interesting insights concerning political electoral results. As a case study we analyze the primary elections of an Italian political party.
State-of-the-art approaches rely on indicators based on tweet and user volumes, often including sentiment analysis. We investigate how to exploit and improve those indicators in order to reduce the bias of the Twitter users sample. We propose novel indicators and a novel content-based method. Furthermore, we study how a machine learning approach can learn correction factors for those indicators. Experimental results on Twitter data support the validity of the proposed methods and their improvement over the state of the art. Moreover we discuss new features coming from Complex Network approach that can be integrated in the predictive model. We believe that predictions based on social network analysis can be significantly improved by exploiting machine learning and complex network tools, where the latter provides valuable high-level features to support the former in learning an accurate prediction function.

3.4 Related work

Many articles propose quantitative approaches to predict the electoral results in different countries: US (OBRS10), Germany (TSSW10), Holland (SB12), Italy (CCP+14). In particular, we distinguish two classes of methods used in literature: volume-based approaches and content-based approaches.
The first class refers to metrics consisting in counting tweets, users, mentions for a given candidate or a political party. (TSSW10) shows that volumes of mentions of parties reflects the distribution of votes in the election among six parties in 2009 German elections. Similar results were achieved by other studies (SB12; BS11). Counting users, instead of tweets, is effective as we can consider each user to be a single elector (DMBR13). Similar approaches were applied to Facebook data as well (Gig12; WG08). For instance, the number of Facebook supporters can be

used as an indicator of electoral success.
Other works highlight some concerns about using tweet volumes to predict elections (SPA$^+$12; MMGA11; GAMM11), showing how in practical cases these approaches may under-perform the baseline. For instance, in (JJS12) it is shown that some arbitrary choices (e.g., the set of considered parties, the time frame, etc.) strongly affect the results, exhibiting a not consistent predictive behavior.
The second class of methods aims at exploiting text information in tweets, and most approaches are based on *sentiment analysis* (BS11). In this context sentiment indicates the degree of agreement expressed in a tweet in relation to a political party or candidate. A few studies applied a machine learning approach to classify tweets according to their polarity, either by training on a manually annotated sample (SB12; BS11) or through dictionary-based unsupervised methods (BS10). Sentiment analysis methods have been used to improve the predictive results of counting methods, but they still are an open research challenge due for instance to the not trivial identification of sarcasm and irony.
Results of both approaches seem not to be consistent across datasets (GAMM11). Predictions vary significantly in relation to the observation period, the data collection and cleansing methods, and the performance evaluation strategy. In fact, all predictive studies have been performed after the outcomes, thus evaluating correlations but not prediction power (MMGA11), and scientific papers are mostly biased towards positive results and they do not report negative ones (Fan10). Finally, the predictive power of Twitter is very sensitive to the bias of its adopters, as Twitter users are not a representative sample of users involved in the elections, neither of people in general. In particular, (SR08) discusses this issue, stating that demographic groups can have different political opinions not equally detectable from new social media. (SB12; GA11) proposed some debiasing strategies.
We adopt as baselines the approach used in (TSSW10; SB12), i.e., counting the mentions of the political candidates in the election, and the one used in (DMBR13), i.e., counting unique users mentioning a candidate. We analyze a data set of tweets related to the 2013 primary elections of the

major Italian political party. The data set is partitioned on the basis of the twenty Italian regions from which the tweets were posted; since we know the electoral results per each region, we can study them as independent election events.
First, we evaluate and discuss state-of-the-art methods based on tweet and user volumes. We, then, propose several new predictors that exploit some enhanced classifications of tweets based on hash-tags. We show that, by properly classifying tweets, it is possible to reduce the error of baseline methods by a factor of 25%.
We also address the bias issue. We propose to learn the degree of bias of each candidate using external polls on expected demographic distribution of voters, so that the prediction can be adjusted accordingly. It turned out that our data set is biased mainly towards young people between 25 and 44 years old and we show that by learning the Twitter bias degree, the electoral ranking outcome can be correctly predicted in 75% of the Italian regions.

3.5 Data

We investigate the echo on Twitter of the primary elections of the Italian major political party: the "Partito Democratico". Our study is conducted on a data set of $\approx$1.7 million tweets. The election took place on December 8th 2013, and the data-set covers about 10 days before and 5 days after the election day. We consider only the geo-located tweets in Italian. In Figure 3 we report a chart with the daily volumes of collected geo-located tweets.

3.5.1 Political context

The "Partito Democratico" is the greatest social-democratic political party in Italy. Three candidate were selected to run for the primary election that took place on December 8th 2013: Mr. Renzi, Mr. Cuperlo, and Mr. Civati. They appeared in the traditional media (TV shows and Press interviews), and they also invested a lot of effort on social media, including Twitter,

in order to create hype and discussions. The candidates received 67.55%, 18.21% and 14.24% of votes, respectively. This result is difficult to predict if we simply base the prediction on Twitter data volumes, because, as shown in the following sections, the presence of Mr. Cuperlo is quite limited compared to the other two candidates. This makes this data set very challenging. Note that Mr. Renzi and Mr. Civati were leading emerging and younger factions in the party.

3.5.2 Data collection and cleansing

The data used in the case study was collected through Twitter API by querying a list of keywords related to the elections and the candidates [1]. The selection of keywords and hash-tags is large enough to guarantee a good coverage of the elections[2].

Data cleansing is a core activity to analyze reliable data. Our initial dataset contained about ≈1.7 million tweets. We deleted partial data and irrelevant tweets provided by Twitter APIs. We selected the Italian tweets on the basis of the language declared by Twitter users and the language detected by a machine learning classifier by Twitter. Only about 8 thousand tweets provided GPS information, whereas the remaining tweets were geo-located by matching the user profile location with the Italian cities and regions.

We finally filtered 95,627 geo-located tweets across the 20 regions of the country, taking into consideration only the tweets published before the election day. The final data set size (≈95 thousand) is comparable with the data sets used in literature, in particular, considering our baseline approaches: namely (TSSW10) where the authors analyzed about 104 thousand tweets covering one month preceding the German elections in 2009, and (DMBR13) where the authors compared different predictive

[1] Data were collected by Michelangelo Puliga, IMT for Advanced Studies. We thank IMT and LinkaLab for the courtesy.

[2] The list of users (through mentions), hash-tags and keywords tracked is the following: *matteorenzi*, *cuperlo*, *civati*, *giannicuperlo*, *vvattuone*, *giannipittella*, *pippocivati*, *giuseppecivati*, *renzi*, *primarie pd*, *partito democratrico*, *primariepd*, *iovotoperch*, *pd*, *matteorisponde*, *congressopd*, *PrimariePD2013*, *cambiaverso*, *pdnetwork*, *ilconfrontopd*,*iostoconcivati*, *ciwati*, *segretario*, *pittella*, *insultacivati*, *d'alema*, *massimoleaderpd*, *dalema*, *giuseppecivati*.

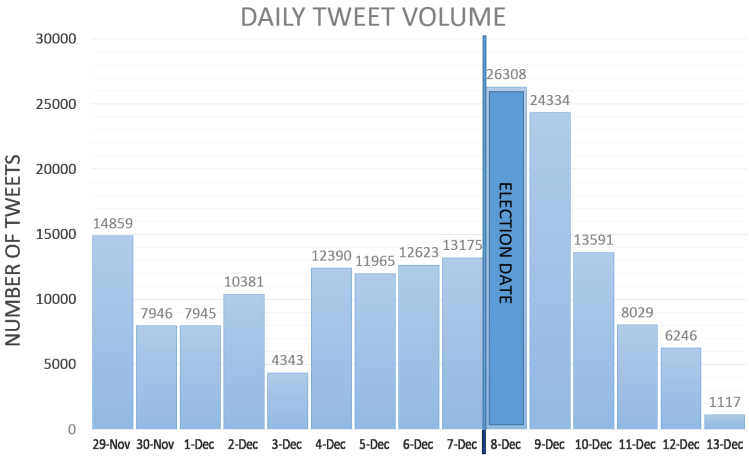


Figure 3: Temporal distribution of tweets

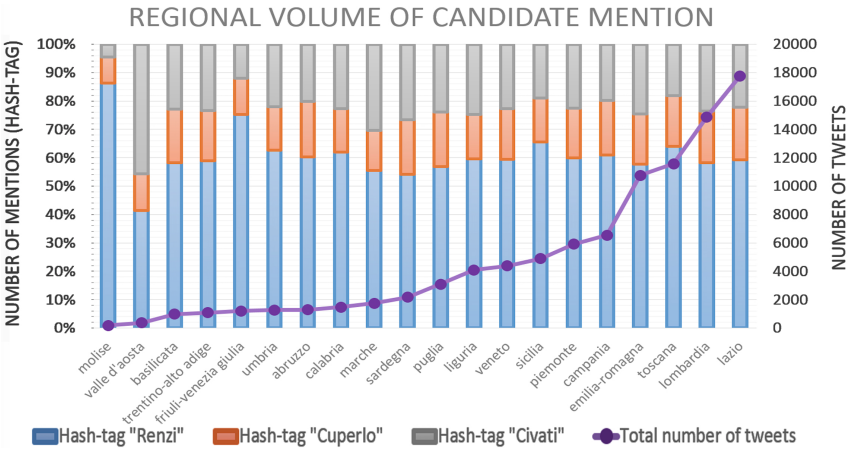


Figure 4: Regional volume of mentions

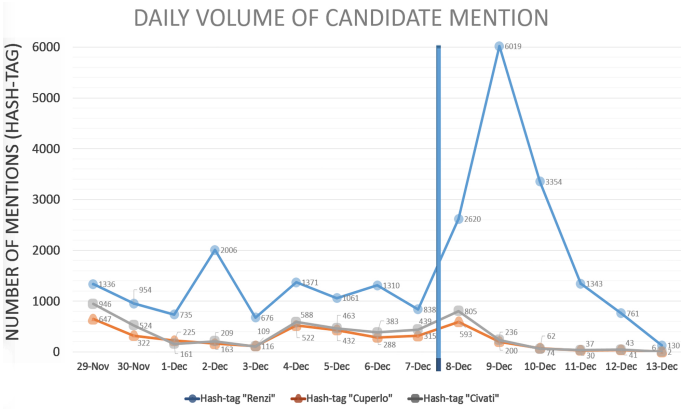


Figure 5: Daily volume of mentions

approaches on a data set of about 114 thousand tweets, covering the three months before U.S. congressional election of 2010.

The time window considered in our experiments is limited to only 10 days before the elections, in line with other works which consider a short time range before the election date being more relevant for predictive tasks, for instance (SB12) (1 week). Figure 4 shows the amount of data collected per region before the election date, and the percentage of mentions of each candidate. Regional volumes are unbalanced and they are correlated to the regional population. The hashtag "renzi" shows a very dynamic nature, with one pick on the 9th December, celebrating the victory of the candidate. Figure 5 shows the distribution of hash-tag occurrences of candidate names over time, before and after the election date.

We investigate those users with the highest posting rate to remove anomalous users. From our evaluation, even the most active users (more than 1 thousand tweets, written in the 10 days before the election) posted meaningful tweets, different from one another, indicating a human behavior. Surprisingly more active users turned out to be individual supporters or local organized groups, not newspapers or official institutional pages.

3.6 Method

In the following we evaluate several *estimators*, or *predictors*. A predictor ϕ produces an estimate $\phi(c)$ of the share of votes that the candidate c will receive. Each predictor is normalized over the set of candidates $\mathcal{C}$ before the evaluation. The normalized version $\bar{\phi}(c)$ is defined as:

$$\bar{\phi}(c) = \frac{\phi(c)}{\sum_{c' \in \mathcal{C}} \phi(c')}$$

We use three different evaluation measures to assess the approaches discussed in this chapter. The most commonly used evaluation measure is the Mean Absolute Error (MAE). We also report the Root Mean Squared Error (RMSE), as it is more sensitive to large estimation errors.
Finally, since we are also interested in the capability of predicting the correct ranking of the candidates, we also introduce the Mean Rank Match (MRM) measure, i.e., the mean number of times that the correct ranking of all the candidates was produced. Note that we conduct a per-region analysis, meaning that a prediction is produced for every region by exploiting the regional data only. The presented results are averaged across the 20 Italian regions.

3.6.1 Baseline

A basic approach is described in (TSSW10). They estimated the share of votes of a political party as the share of tweets mentioning it. Let $\mathcal{T}$ be set of tweets in the observed period, and let $\mathcal{C}$ be the set of parties, the popularity $f(c)$ of a party is defined as:

$$\phi(c) = f(c) = |\{t \in \mathcal{T} \mid c \in t\}|$$

where $c \in t$ holds *iff* the tweet t mentions the party c (in our case study we consider different candidates in a primary election, which are assimilated to parties running in a political election). Understanding whether a tweet discusses a given political party may not be straightforward.
In (TSSW10), a tweet is considered to mention a given political party if its text contains the party acronym or the name of selected politicians of

Table 8: Baseline methods performance.

Algorithm	MAE	RMSE	MRM
TweetCount	0.0818	0.1024	0.35
UserCount	0.0940	0.1080	**0.45**

Table 9: Classification methods performance: MAE, RMSE, relative difference of RMSE over the baseline with the lower RMSE, MRM.

Algorithm	MAE	RMSE	Δ	MRM
UserShare	0.0616	0.0792	-22.7%	0.35
ClassTweetCount$_{\mathcal{H}}$	0.1056	0.1248	+21.9%	0.30
ClassUserCount$_{\mathcal{H}}$	0.0924	0.1090	+6.4%	0.30
ClassTweetCount$_{\mathcal{C}}$	**0.0636**	**0.0786**	-23.2%	0.35
ClassUserCount$_{\mathcal{C}}$	0.0804	0.1033	+0.9%	0.40

the party. This simple estimator achieves a MAE of 1.65% and it was able to predict the correct ranking of the elections. Authors conclude that $f(c)$ can be used as a plausible estimation of vote shares, and they show that this estimator is very close to traditional election polls.

Users counts, instead of tweet counts, are considered in (DMBR13). Let $\mathcal{U}$ be the set of twitter users, the popularity $u(c)$ of a party is defined as the number of users mentioning c at least once in the observed period:

$$\phi(c) = u(c) = |\{u \in \mathcal{U} \mid \exists t_u \in \mathcal{T} \wedge c \in t_u\}|$$

where t_u denotes a tweet t authored by user u. The $u(c)$ predictor showed to be only marginally better. We named the above two methods TweetCount and UserCount respectively.

In our analysis, we considered a tweet to mention a candidate if it contains a hash-tag with his family name, i.e., `#renzi`, `#cuperlo` or `#civati`. The performance measures on our data set are reported in Table 8. The performance of the first two methods are very close both in terms of MAE and RMSE.

We can observe some improvement in terms of MRM, suggesting the focusing on Twitter users as estimators of the behavior of voters is a valuable approach.

Considering the full text instead of hash-tags with these predictors did not provide any significant benefit, and therefore results are not reported here. We exploit the full text in some content-based predictors presented later.

3.6.2 Exploiting tweet/user classification

We first propose an improvement over the UserCount strategy. According to UserCount, the relation according to which a Twitter user corresponds to one voter is not satisfied as users mentioning more than one candidate are taken into consideration multiple times.
We correct this behavior with a normalization by the number of candidates mentioned. We say that a user $u \in \mathcal{U}$ is likely to vote for candidate $c \in \mathcal{C}$ with probability $P(c|u)$, defined as:

$$P(c|u) = \frac{\mathbb{1}\{\exists t_u \in \mathcal{T} \wedge c \in t_u\}}{|\{c' \in \mathcal{C} | \exists t_u \in \mathcal{T} \wedge c' \in t_u\}|}$$

where $\mathbb{1}\{x\}$ is equal to 1 if x is true and 0 otherwise. Clearly, $\forall u \in \mathcal{U}, \sum_{c \in \mathcal{C}} P(c|u) = 1$. We thus estimate the number of users likely to vote candidate c as:

$$\mathsf{UserShare}(c) = \sum_{u \in \mathcal{U}} P(c|u)$$

In the following we propose some enhanced classification of tweets polarity for the candidates.
We try to evaluate what is the probability that mentioning a hash-tag h leads to a vote for a given candidate c. We introduce an approximation here, with the usual assumption that mentioning a candidate is equivalent to voting a candidate. Then, we can easily estimate $P(c|h)$ as follows:

$$P(c|h) = \frac{P(c,h)}{P(h)} = \frac{|\{t' \in \mathcal{T} | c \in t' \wedge h \in t'\}|}{|\{t' \in \mathcal{T} | h \in t'\}|}$$

This has the effect of smoothing the impact of very frequent hash-tags which are likely to occur frequently with every candidate mention, thus not providing any significant signal. By focusing on the subset of the

100 most frequent hash-tags $\mathcal{H}$, each tweet $t \in \mathcal{T}$ is associated with a candidate $c \in \mathcal{C}$ according to the score:

$$S_{\mathcal{H}}(c|t) = \sum_{h \in t \cap \mathcal{H}} P(c|h)$$

According to $S_{\mathcal{H}}(c|t)$ every hash-tag in t may contribute to strengthen the relation with a given candidate $c \in \mathcal{C}$. We can now use $S_{\mathcal{H}}(c|t)$ to *label* a tweet with a candidate. We say that t is labeled with c, or equivalently $\lambda_{\mathcal{H}}(t) = c$, if $c = \arg\max_{c' \in \mathcal{C}} S_{\mathcal{H}}(c'|t)$. Whenever $\lambda_{\mathcal{H}}(t)$ is non uniquely defined, i.e., multiple candidates have the same score, t is assigned to c with probability $\bar{f}(c)$, where $\bar{f}(c)$ is the normalized tweet count. We finally introduce a new indicator measuring the count of tweets labeled with a given candidate:

$$\mathsf{ClassTweetCount}_{\mathcal{H}}(c) = |\{t \in \mathcal{T} \mid c = \lambda_{\mathcal{H}}(t)\}|$$

This indicator is extended to consider users rather than tweets. We say that u is labeled with c, or equivalently $\lambda_{\mathcal{H}}(u) = c$, if $c = \arg\max_{c' \in \mathcal{C}} |\{t_u \in \mathcal{T} | c' = \lambda_{\mathcal{H}}(t_u)\}|$. Whenever $\lambda_{\mathcal{H}}(u)$ is non uniquely defined, i.e., multiple candidates have the same score, u is assigned to c with probability $\bar{f}(c)$. We therefore define an indicator counting the number of users labeled with a given candidate:

$$\mathsf{ClassUserCount}_{\mathcal{H}}(c) = |\{u \in \mathcal{U} \mid c = \lambda_{\mathcal{H}}(u)\}|$$

We finally found interesting to focus on the candidates mentions only instead of the set of hash-tags $\mathcal{H}$. Analogously to $\mathsf{ClassTweetCount}_{\mathcal{H}}$ and $\mathsf{ClassUserCount}_{\mathcal{H}}$, we can define new labeling functions $\lambda_{\mathcal{C}}$ based on a new score function $S_{\mathcal{C}}$:

$$S_{\mathcal{C}}(c|t) = \sum_{h \in t \cap \mathcal{C}} P(c|h)$$

Given $\lambda_{\mathcal{C}}$, we thus define the following strategies:

$$\begin{aligned} \mathsf{ClassTweetCount}_{\mathcal{C}}(c) &= |\{t \in \mathcal{T} \mid c = \lambda_{\mathcal{C}}(t)\}| \\ \mathsf{ClassUserCount}_{\mathcal{C}}(c) &= |\{u \in \mathcal{U} \mid c = \lambda_{\mathcal{C}}(u)\}| \end{aligned}$$

Table 9 shows the performance of the above strategies exploiting classification of tweets and users. The two most promising are UserShare and ClassTweetCount$_{\mathcal{C}}$. These strategies are both very simple as they consider only the hash-tags corresponding to candidates mentions.
In UserShare, a single user vote is *split* among the candidates, while in the strategy ClassTweetCount$_{\mathcal{C}}$ a tweet is classified as a vote to only one of the candidates. Both approaches provide a significant improvement of about 25% over the baseline strategies both in terms of MAE and RMSE. The MRM score is still too low to draw final conclusions.

3.6.3 Training correcting factors

One of the assumptions of the we present here is that Twitter users are not a representative sample of the voters population. Even if we were able to correctly classify each Twitter user, we would not be able to make a reliable estimate of the voting results as (i) several Twitter users may not vote, (ii) several voters are not present on Twitter, (iii) several Twitter users may not express their political preferences or they may alter them and (iv) the voters of each candidate have a different degree of representativeness in Twitter.
Given a predictor $\phi(c)$, we aim at learning a set of weights w_c, one for each candidate, such that $w_c\phi(c)$ improves the estimate of actual votes received. The weights w_c should act as a bridge correcting an estimate based on Twitter users to fit real world users behavior.
We aim at *learning* the weights w_c. For each region of Italy and for each candidate c, we create a training instance $\langle y_c, x_c \rangle$, where y_c is the *target variable* being equal to the percentage of votes actually achieved by c in the given region, and x_c is the *input variable* equal to a given estimator $\phi(c)$.
In general, a vector of *input variables* can be used. We thus have a *training data set* with 60 training instances coming from 20 regions and 3 candidates. To conduct a 5-fold cross validation the data set are split region-wise in training and test sets. The training set is used to learn a weight w_c via linear regression that minimizes $(y_c - w_c \cdot \phi(c))^2$. We apply this

Table 10: Machine-learned weighting performance

Algorithm	MAE	RMSE	MRM
ML-UserShare	0.0536	0.0705	**0.75**
ML-ClassTweetCount$_{\mathcal{C}}$	0.0533	0.0663	0.69
ContentAnalysis	**0.0525**	**0.0630**	0.70

approach to the two most performing predictors evaluated so far, i.e., UserShare and ClassTweetCount$_{\mathcal{C}}$. We name the corresponding *machine learned* strategies ML-UserShare and ML-ClassTweetCount$_{\mathcal{C}}$. As reported in Table 10 these new approaches provide a significant improvement according to all metrics. The improvement is of about 15% in terms of MAE and 10% in RMSE.

A huge improvement is observed according to the MRM metric. For instance, ML-UserShare is able to provide the correct candidate ranking in 15 out of 20 regions. This means that we are able to reduce the prediction error on the votes share (both MAE and RMSE) up to the point of being able to correctly predict the final ranking of the candidates.

By inspecting the weights learned by the ML-UserShare strategy, we see that Renzi, Cuperlo and Civati have weights 1.02, 1.24 and 0.70 respectively. This means that the second candidate is *under-represented* in the Twitter data, and symmetrically for the third candidate.

In Figure 6 we show the actual voting results and the estimations produced by UserShare and ML-UserShare. The correcting weights of ML-UserShare have sometimes the effect of inverting the rank generated by UserShare of the two candidates Cuperlo and Civati, in agreement with the actual election results.

The drawback of this approach is that it requires a training data where to learn the correction weights w_c. This makes it not possible to directly apply the method before the election takes place.

On the other hand, we can assume that weights are sufficiently stable, i.e., that the degree of representativeness of the Twitter sample for a specific sample does not change abruptly. If this is the case, then we can learn those weights by exploiting data from previous events.

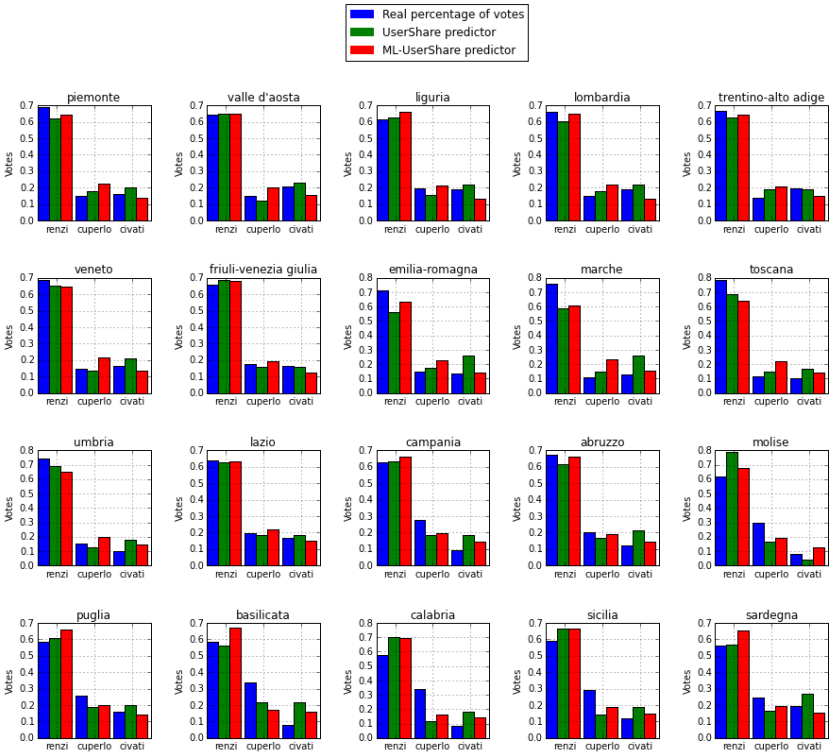


Figure 6: Regional predictions and actual voting results

Indeed, it would be possible to exploit elections at municipality, regional and European level to learn a proper set of weights for national elections. Another interesting case is that of a two-round voting system, where the model could be trained after the first round and used to predict the outcome of the second.
Yet another option is to complement prediction with traditional polls data.

3.6.4 Including content-based analysis

The above approach is very general as several features about a candidate can be considered altogether by extending the input variable x to a vector of input variables. We propose to include text analysis and semantic analysis as follows. We consider the top 100 words (not only hash-tags) most frequently occurring in the data set after stop-word removal and stemming. These 100 words $\mathcal{W}$ are used to build a *content-based feature vector*.

For each candidate and for each region, we compute the number of occurrences of each word in $\mathcal{W}$ normalized by the number of tweets in the region considered. This new feature vector include the names of the candidates, but it is also likely to include, if frequent, other significant *names*, *topics* or *catch-phrases* which are relevant to estimate the reach of a candidate.

Similarly as for ML-UserShare, we build a new training data set where for each training instance $\langle y_c, x_c \rangle$, x_c contains the *content-based feature vector* for c, to which we also include the predictor UserShare. The resulting model should be able to blend text analysis with the predictor UserShare. The weight vectors for each candidate are learned with LASSO linear regression. The resulting predictor is named ContentAnalysis. As shown in Table 10, ContentAnalysis achieves the best MAE and RMSE, and a good value of MRM.

3.6.5 Demographic analysis

We think that the main issue of any social network analysis, aiming at understanding public opinion, is that social networks are not a representative sample of people, or, in this context, of the voters. The bias introduced by Twitter should be carefully taken into consideration. From the data we collected, it is not possible to infer details about users, e.g., age, education or other. We resort to analyze the Twitter demo-graphical bias through external polls on the age distribution of voters[3]. We compare the UserShare predictor against the expected result of 5 age range classes.

[3]Data from polls performed by Quorum (polling Institute).

Table 11: Error of UserShare by age class

Age class	MAE
16-24 years	0.1409
25-44 years	0.0216
45-54 years	0.0476
55-64 years	0.0636
$>$ 65 years	0.0709

Table 12: Estimations at national level

Algorithm	MAE	RMSE
TweetCount	0.0541	0.0641
UserShare	0.0413	0.0462
Polls	0.0386	0.0418

The results are reported in Table 11 ordered by MAE, showing that UserShare is more accurate in predicting the votes of people in the range of 25-44 years old. It is known that the average age of Italian Twitter users is 32 years (larger than the world average age which is 24), according to a report of Pew Research published in 2013, confirming our preliminary results. This suggests that Twitter analyses and traditional polls can be complemented together in order to achieve a wider coverage.

3.6.6 Aggregated outcome

Finally, in order to provide a full picture of our analysis, we provide estimations at national level, i.e., by considering the whole data-set without partitioning by region and without using machine learning methods.
Table 12 shows the performance of TweetCount (TSSW10) and UserShare. We also report the average error of the electoral polls made by different polling institutes (period 26 Nov - 04 Dec), as it is reported in *termometropolitico.it*, a website which collects and comments political polls before elections.
The two methods TweetCount and UserShare are very close to the polls error, and we can explain this error with the age sampling bias which is discussed in the previous section. Note that we don't use any machine learning to improve the prediction in this case.
Finally, recall that the cost of traditional polling is obviously higher that the cost of twitter monitoring.

3.6.7 Beyond counting tweets

The naïve approach of correlating simple social media networks measures, e.g., tweets volume, is not often sufficient to provide accurate estimation of real world phenomena.
We believe that machine learning methods are capable of devising more accurate models, by exploiting social media features in a non trivial way. We aim at exploiting network properties to support machine learning algorithms.
The application of machine learning methods is harmed by the lack of positive training instances, e.g., elections are not very frequent. Therefore, we need machine learning methods able to generalize well and minimize mis-prediction risk with a very small number of positive examples.
The dynamism of social network data and their size require new network analysis tools that take into account the network evolution and that provide accurate methods of streaming analysis.

3.7 Conclusion

In this chapter, we tackled the problem of providing accurate estimation of real world phenomena through polarization analysis with three novel contributions in the context of vote prediction.
First, we evaluated counting-based state-of-the-art methods, and we proposed an enhanced user centered predictor that models every single user with a voting probability across the candidates. This predictor improves by 25% the baseline methods.
Then, we addressed the main issue of the social network sample bias. We proposed a few machine learning approaches, also including content-based analysis, with the goals of learning bias correcting factors. In our case, we were able to estimate the over or under representativeness of each candidate in our data, but the cross-validation method can be used only retrospectively. We believe that exploiting machine learning, both for an improved classification of users and for correcting the sample bias is a crucial task in social network analysis. The main drawback of

such techniques is that they require training data. We believe that such drawback can be overcome by exploiting continuous analysis over time leveraging related events, e.g., political elections at any level. How to transfer the knowledge gained in one analysis to other scenarios is still an open research problem.

In conclusion, we believe that the analysis of *polarized communities* is OSNs can be used to predict collective social behavior, but major improvements in the field can be achieved by integrating several sources of information, such as traditional polls, multiple social networks, demographic data, historical data, analyses of related events, content-based and network-based properties. Such wealth of information can be exploited altogether through machine learning approaches. The integration of all of these approaches may open up new research challenges and opportunities in the field.

Chapter 4

Analytical framework: time, places, and polarization

The results discussed in this chapter were published in (CLM+16; CLOP16).

- Coletto, M., Esuli, A., Lucchese, C., Muntean, C. I., Nardini, F. M., Perego, R, Renso, C. (2016). Sentiment-enhanced Multidimensional Analysis of Online Social Networks: Perception of the Mediterranean Refugees Crisis. In Workshop on Social Network Analysis Surveillance Technologies, co-located with IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM 2016, August 18-21, San Francisco, CA, US.
- Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2016). Polarized User and Topic Tracking in Twitter. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 945-948). SIGIR 2016, July 17-21, Pisa, Italy.

4.1 Introduction

Predicting social behavior with data coming from OSNs is a challenging task mainly because Social Media are not an unbiased sample of the

population and because of the large number of variables involved in the human decision process that should be considered in a model.
However, OSNs contain much more information that can be exploited in addition to *polarization*. By exploiting this additional information we can dig more into the perceptions of the users, which is the preliminary step to subsequently predict their behavior.
Content of the messages exchanges and metadata, for instance, provide indications about locations and time. Then, *polarization*, or more simply the *sentiment* of a user, can be integrated with additional information to create a framework of analysis of complex social phenomena in OSNs.

4.2 Problem formulation

In this chapter we aim at answering the following research question:
Q3: *Can we integrate polarization with other variables (i.e., time, space) to create an analytical framework that might be used to study social phenomena?*
The current chapter is dedicated to the integration of polarization with other variables in order to create a framework that can be used to analyze a social phenomenon. We use the framework to analyze the recent issue of Mediterranean refugees and the perception of the phenomenon by European countries in Twitter.

4.3 Contribution

We propose an analytical framework able to investigate discussions about polarized topics in online social networks from different angles. The framework supports the analysis of social networks along several dimensions: time, space and sentiment.
We use the notion of sentiment which is a subset of polarization as we discuss in Chapter 1 because it is more simple to detect in our case study. The algorithm used to detect the sentiment is the same proposed in Chapter 2 to detect polarization, but the initial input in terms of keywords are two simple sets of terms pro and against the social phenomenon studied, representing the positive and the negative sentiment. Of course the same

framework could integrate additional polarization classes without loss of generality. In the application that we describe in this chapter we simplify the situation by considering only two opposing polarization classes and then we refer to polarization in terms of sentiment.
In this chapter we show that the proposed analytical framework can be used to interpret social trends from large tweet collections by extracting and crossing information about the following three dimensions: time, location and sentiment. We describe the methodology to: 1) extract relevant spatial information, 2) enrich data with the sentiment of the message and of the user (retrieved in an automatic iterative way), 3) perform multidimensional analyses considering content and locations in time. The approach is general and can be easily adapted to any topic of interest involving multiple dimensions.
For the scope of chapter we use our framework to outline the European perception of the refugee crisis. Our study shows differences in positive and negative sentiment in EU countries, in particular in UK, and by matching events, locations and perception, it underlines opinion dynamics and common prejudices regarding the refugees.

4.4 Related work

We are recently witnessing one of the largest movement of migrants and refugees from Asian, African and Middle-east countries towards Europe. The United Nations High Commissioner for Refugees (UNHCR) estimates one million of refugees arrived to the Mediterranean coasts in 2015 mainly from Syria (49%), Afghanistan (21%) and Iraq (8%). Figure 7 reports main routes to EU coasts and to northern Europe.
The largest wave of arrivals started in August 2015 following a main route through Turkey, Greece, Macedonia, Hungary and Austria to Germany, France, UK and other northern European countries. Since then, this phenomenon has been in spotlight of the media, which have reported an increasing number of events related to migrants, such as the additional border controls established by Hungary, Austria and Germany, the several incidents involving refugees, or the story of the young Syrian boy found

dead on the seashore in Turkey in September 2015.
The implications of this refugee crisis are complex. The whole phenomenon is nowadays object of a heated and polarized debate. Understanding how the debate is framed between governmental organizations, media and citizens may help to better handle this emergency.
Through the analysis of the Twitter online social network, we address the following questions: "How is the European population perceiving this phenomenon? What is the general opinion of each country? How is perception influenced by events? What is the impact on public opinion of news related to refugees? How does perception evolve in time in different European countries?" Social media may help in answering these relevant questions but the volume of messages exchanged is massive and the extraction of sentiment is challenging.
Basic analyses of the phenomenon through Twitter have been already

Figure 7: The routes to European countries - source Business Insider from Europol, Reuters, Washington Post, AFP, ICMPD.

performed by media in a simplistic way, mainly through manual analysis of content, news and hashtags for small samples of tweets. For instance, it has been found that the *#welcome refugees* and *#germany* hashtags are used mainly from outside Germany or in other articles the usage of terms *refugees* and *migrants* have been compared in offline and online media: the former indicates someone forced to leave her country to avoid war or imprisonment, the latter is instead someone moving from his/her country searching for better living conditions. These are few examples of simple analyses on this phenomenon [1]. However a more comprehensive work trying to analyze the perception of users about these events, shaped across places and time, is still missing.

There is a significant increase of interest in collecting and analysing geo-located data from online social networks (OSNs). Several works study different aspects of the geographical dimension of OSNs, a broad study on this argument is reported in (SMML10). The authors propose a framework to compare social networks based on two new measures: one captures the geographical closeness of a node with its network neighborhood and a clustering coefficient weighted on the geographical distance between nodes.

Twitter geo-located posts are studied in (TGW12) to understand how Twitter social ties are affected by distance. Linked users are identified as "egos" and "alters" and the distance between them is analyzed by considering the correlation with the air travel connection distance and with national borders and languages.

An analogous objective is the focus of (KKNG12) where the authors infer the location of 12 million Twitter users in a world-wide dataset. Differently from the previous paper, they study the correlation between the Twitter population and the socio-economic status of a country, suggesting that highly developed countries are characterized by a larger Twitter usage.

The geographical properties of Twitter are also useful to study the movements of people and migration phenomena.

A study of mobility using geo-located Twitter messages is presented in

[1] `http://orientalreview.org/2015/09/21/`
`https://storify.com/ImagineEurope/what-is-associated-with-europe/`

(HSB+14). The authors introduce a detailed study aimed at estimating international travelers based on the country of residence. They identify a number of characteristics including radius of gyration and mobility rate to describe the traveling phenomena thorough the Twitter lens.
Authors in (ZGWS14) show how the analysis of 500,000 geo-located Twitter users may help to predict the migration turning points and to better understand migration in OECD countries. The authors estimate the migration rate of users moving from one "home" country to another country. The reported results depict some interesting trends such as the decrease of migration from Mexico to US, consistent with official estimations.
Twitter is also exploited to better understand how the communication flows during political movements and events (CDF+13). This work studies Twitter data covering the birth and maturation of the American anti-capitalist movement *Occupy Wall Street*. The authors analyze the geo-spatial dimension of tweets in combination with the communication dimension building a geographic profile for the communication activity of the movement. An extensive analysis of these data produced many interesting results. For example, it appears that proximity to events plays a major role in determining which content receives the most attention in contrast to the stream of domestic political communication.
As we have already deeply discussed in the previous chapter, using Twitter for opinion mining and user polarization is a vast subject (PL08). The sentiment analysis methods proposed are many, mainly based on dictionaries and on learning techniques through unsupervised (PP10) and supervised methods (lexicon-based method (TBT+11)) and combinations (KSTA15).
Opinion mining techniques are widely used in particular in the political context (AG05b) and in particular on Twitter (CLOP15).
Recently new approaches based on polarization, controversy and topic tracking in time have been proposed (GDFMGM16). Among them the method described in Chapter 2 is built on the evidence that polarized users in an OSN are grouped based on their opinion on a particular topic. We use this approach in this chapter to study the social phenomenon of

Table 13: Notation

Symbol	Description	# Total
$\mathcal{G}$	Collected English tweets	97,693,321
$\mathcal{T}$	Tweets related to the refugee crisis	1,238,921
$\mathcal{T}_{c+}$	Positive sentiment tweets	459,544
$\mathcal{T}_{c-}$	Negative sentiment tweets	387,374
$\mathcal{T}_{ML}$	Tweets with mentioned location	421,512
$\mathcal{T}_{UL}$	Tweets with user location	101,765
$\mathcal{U}$	Users	480,660
$\mathcal{U}_{c+}$	Users with positive sentiment	213,920
$\mathcal{U}_{c-}$	Users with negative sentiment	104,126
$\mathcal{U}_{L}$	Users with country location	47,824

refugees. All these effective approaches are based on network measures and clustering (GDFMGM16) or hashtag classification through probabilistic models (CLOP16) with no use of dictionary-based techniques, which have many limitations due to the uncertainty of natural language.
The novelty of our proposal compared to the state-of-the-art approaches is mainly the fact that we introduce an analytical framework to study a mass event from Twitter messages as a combination of three dimensions: time, space and sentiment. The sentiment analysis method adopted is efficient in tracking polarization over Twitter w.r.t. other more generic methods. Differently from many approaches studying migration, we do not base our analyses on the change of location of Twitter users to measure the flow of individuals through space, but rather we aim at understanding the impact on the EU citizens perception of migrants' movements.

4.5 Data

In this section we detail the data collection phase and the analytical dimensions, namely the spatial, temporal and sentiment dimension. The final multi-dimensional dataset can be analyzed and queried along these axes and, more interestingly, on combinations between them. The data statistics and the notation used are summarized in Table 13.
We use the Twitter Streaming API to collect English tweets data under the *Gardenhose* agreement (10% of all tweets in Twitter) in period from mid

August to mid Sept 2015, noted with $\mathcal{G}$, out of which we selected the tweets related to the refugee crisis topic, called the *relevant* tweets (denoted as $\mathcal{T}$). We did this by manually choosing a subset of 200 hashtags frequently used and specifically related to refugees in the period of analysis.
From $\mathcal{T}$ we extract information about three main dimensions: **spatial, temporal, and sentiment-based**, resulting in the set of users and tweets as reported in Table 13.

4.5.1 Spatial and temporal dimensions

For each tweet we extract two kinds of spatial information if present: the *user location* of the person posting the message and the *mentioned locations* within the tweet text. The *user location* is structured in two levels, the city (if present) and the country. The user city is identified from the *GPS coordinates* or *place* field when available. Since GPS and *place* data are quite rare (about 3.5K in all dataset) we used the free-text *user location* field to enrich location metadata.
We identified locations in the user generated field based on location data from the Geonames[2] dictionary which fed a parsing and matching heuristic procedure. This technique provides high-resolution, high-quality geolocation in presence of meaningful user location data (OAG$^+$11). The user country is collected in a similar way and when not explicitly present we infer from the city. The *mentioned locations* in the text are also represented at city and country level, and they are extracted from tweets' text with the same heuristic procedure as for user location. We limit our analysis to the perception and sentiment of European citizens. For the mentioned locations we are also interested to the countries involved in the migration crisis. The numbers of tweets with user location $\mathcal{T}_{UL}$ and mentioned locations $\mathcal{T}_{ML}$ are reported in Table 13.
Finally, we extract the publishing time of each tweet and the period of time when each user was active. This information is necessary to study the evolution of the migrant crisis phenomenon and users' perception over time.

[2]http://www.geonames.org/

4.5.2 Sentiment dimension

We are interested in understanding if the user has a positive feeling in welcoming the migrants or if he/she mainly expresses negative feelings (fear, worry, hate). Therefore, the dataset is enriched with information about the sentiment for both of tweets and users.
We consider two polarization classes $c \in \mathcal{C}$: *pro refugees* (c_+) and *against refugees* (c_-). We apply the algorithm PTR (Polarization Tracker) (CLOP16) to assign a class to each polarized tweet and to each polarized user in an iterative way as we deeply described in Chapter 2. The approach proposed in (CLOP16) is suitable to track polarized users according to a specific topic which is in our case the "refugees phenomenon". The initial seeds have been selected by analyzing the most frequent among about 95 K unique hashtags:

$H^0_{c_+}$ = `#refugeeswelcome #refugeesnotmigrants #welcomerefugees`

$H^0_{c_-}$ = `#refugeesnotwelcome #migrantsnotwelcome #norefugees`

The initial seed $H^0_{c_+}$ is used in 36K tweets, whereas $H^0_{c_-}$ hashtags are used in only 2K tweets. One of the benefits of PTR is that after only a few iterations the results are less dependent on the size of the original seed, correcting the unbalanced number of occurrences per class.
The procedure adds information about polarization of the users by polarized hashtags extension through the analysis of all the tweets written by an already polarized users and not only the polarized tweets. The iterative procedure reaches the convergence after 4 iterations. We exclude from the hashtags retrieved by PTR all the hashtags which directly mention a city or a country. This is to keep the sentiment value independent by the location in the computation of the polarization. The combination of location and sentiment is done by crossing the space and sentiment.
Alg. 1 extracts new hashtags at each iteration. We report the most relevant retrieved hashtags in addition to seed ones after the final iteration of the algorithm for each class c:

$H^{\tau=final}_{c_+}$: `#campliberty #health #humanrights #marchofhope #migrantmarch #refugee #refugeecrisis #refugeemarch #refugeescrisis #sharehumanity #solidarity #syriacrisis #trainofhope`

$H_{c\_}^{\tau=final}$: #alqaeda #guns #illegalimmigration #illegals #invasion #isis #islamicstate #justice #migrant #migrantcrisis #muslimcrimes #muslims #no2eu #noamnesty #nomoremigrants #nomorerefugees #patriot #quran #stoptheeu #taliban #terrorism

From the analysis of the extracted hashtags we can see that people with a positive sentiment prefer to use the term *refugees*, while people with a negative sentiment refer to them as *migrants*, thus minimizing the fact that they are escaping war and persecution. Users with a negative sentiment frequently use *refugees* and the Islamic religion together, somehow correlating, in a prejudicial way, refugees with Islam and terrorism. Finally, we observe that individuals with negative sentiment are often patriotic and not pro Europe.

Note that the algorithm, may classify both users and tweets as non polarized, thus favoring accuracy of truly polarized content. The polarization algorithm was able to assign the sentiment to 68% of the tweets and to 66% of the users in our dataset. Regarding EU-geolocated tweets and users, the algorithm assigned the sentiment to 73% of tweets and to 71% of the users. The sentiment analysis has been performed through PTR (CLOP16) since this method does not need external dictionaries or supervision and provides a classification of polarized users in a flexible way by looking at terms used by members of different opinions. In our case the method suits our task to study polarization of Twitter users in relation to the topic of refugees.

4.6 Analytical framework

Our study is driven by the analytical questions below:

AQ1: *What is the evolution of the discussions about refugees migration in Twitter?*

AQ2: *What is the sentiment of users across Europe in relation to the refugee crisis? What is the evolution of the perception in countries affected by the phenomenon?*

AQ3: *Are users more polarized in countries most impacted by the migration flow?*

4.6.1 Spatial and temporal analysis

We explore our multidimensional dataset by first analysing the spatial and temporal dimensions to answer AQ1. These analysis quantifies the volumes of relevant Twitter messages based on the countries of the users and the country mentions, since these volumes are strong indicators of real-world events (WL11).
Figure 8 depicts the total number of tweets for the 20 most active countries.

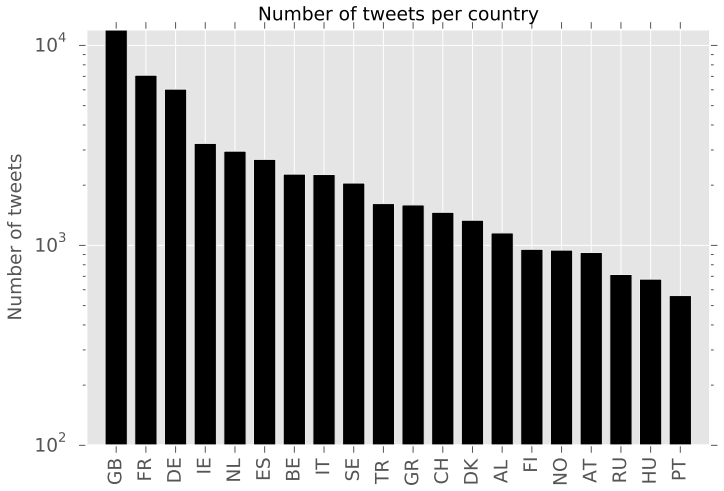


Figure 8: $\mathcal{T}_{UL}$ per top-20 countries in log scale

Since the dataset is in English most of the tweets (56.1%) come from users located in United Kingdom (UK), therefore in Section 4.6.2 we focus our analysis on UK. Nevertheless, a significant fraction of the data come from other countries, e.g., France (FR) accounts for 6.9% of the tweets and Germany (DE) accounts for 5.9%. Without loss of generality, our methodology can be extended to other languages by simply extending the seed hashtags used in the sentiment dimension construction. As far as the mention location is concerned, we see that users from 51 countries

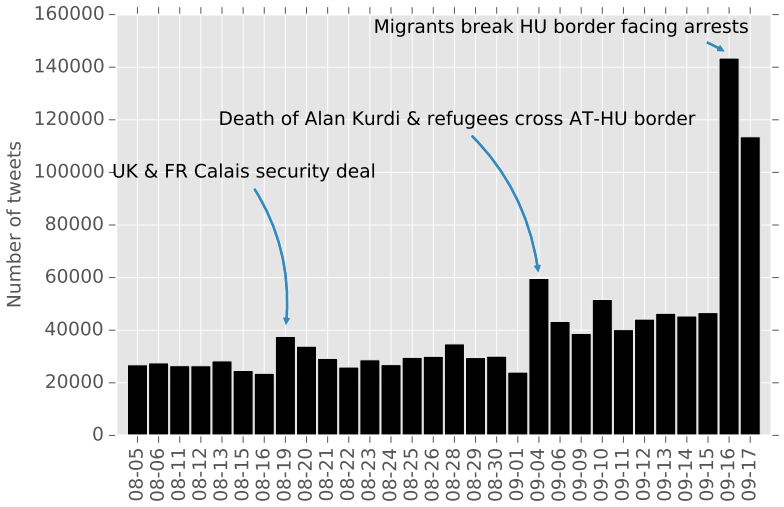


Figure 9: $\mathcal{T}$ per day and top pieces of news

mention 154 countries in Europe, Asia and Africa. This is analyzed in further detail in Section 4.6.2.

Figure 9 illustrates tweets volumes along the temporal dimension, therefore relating volumes to events, summarized in Table 14. This clearly includes all the tweets in $\mathcal{T}$ and not only the geo-located ones. We observed significant volume peaks in days August 19, September 4 and September 19. As we can see from the table, these days match the major events since the UK and France security deal signed on the 18 of August regarding Calais, the drowned Syrian boy found on the beach in Greece, Hungary takes refugees to Austrian border by bus in days 2 to 4 September, and migrants breaking through Hungarian border on September 16.

Next, we analyze location mention to a country related to the refugee migration.

Table 14: Major events reported by UK newspapers - Events happened during the observation period.

18.08	UK and France to sign Calais security deal.
20-21.08	Macedonian police teargas thousands of refugees crossing from Greece and declares state of emergency over surge in migrants & refugees.
27-28.08	71 dead refugees found dead in truck in Austria.
31.08	Angela Merkel: Europe as a whole must help with refugees.
1.09	Hungary closes main Budapest station to refugees.
2.09	Alan Kurdi drowned off the shores of Turkey.
4-6.09	Migrants are allowed to cross the Austro-Hungarian border; Refugees welcomed warmly in Germany.
8.09	Hungarian Journalist appears to kick and trip fleeing refugees.
14.09	Austria followed Germany's suit and instituted border controls; Refugee boat sinking: dozens including children drown off Greek island.
15.09	Croatia started to experience the first major waves of refugees; Hungary announced it would start arresting people crossing the border illegally.
16.09	Refugee crisis escalates as people break through Hungarian border; Hungary had detained 519 people and pressed criminal charges against 46 for trespassing, leading to pursue alternative routes through Croatia from Serbia.
17.09	Croatia decided to close its border with Serbia.

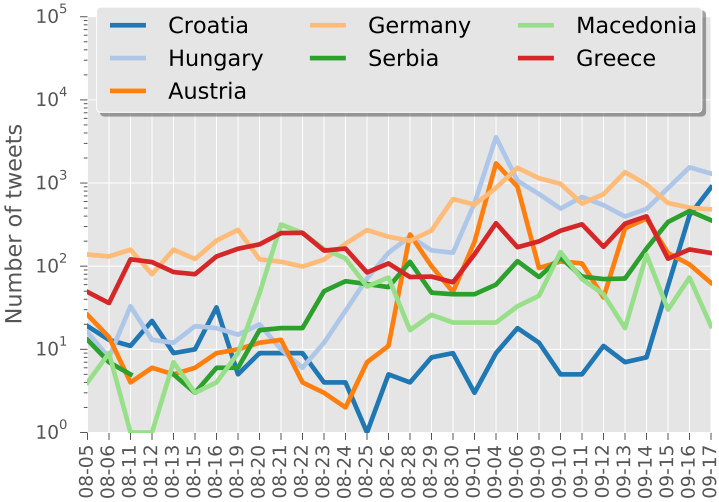


Figure 10: EU country mentions per day in log scale

Figure 10 reports the volumes of tweets mentioning the EU countries most impacted by the refugees route, namely Austria, Germany, Croatia, Macedonia, Hungary, Serbia, Greece.
We see that there is an interesting correspondence between the peaks of mentions and the events timeline. An evident peak for Germany, Austria and Hungary is the first week of September, probably related to the news of borders being opened to refugees. We also notice a peak of mentions of Croatia corresponding to the closing of borders with Serbia.
Macedonia also sees an important increase of mentions around the 20th of August, probably in relation to the Macedonian Police using tear gas on refugees.
Similarly, Figure 11 focuses on the mentions of relevant non European countries. The number of tweets mentioning Syria increases dramatically after the aforementioned facts of September 4, Turkey has a peak the day 4 of September due probably to the Alan Kurdi news. We also observe how the mentions to other countries remain more or less stable along this

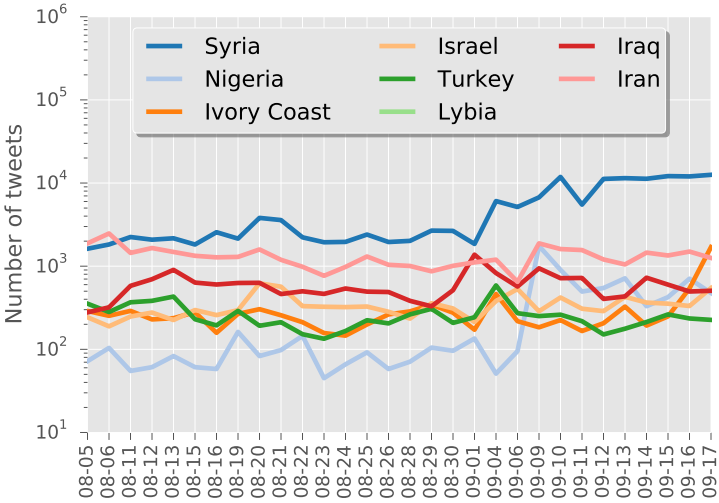


Figure 11: Non-EU country mentions per day in log scale

period to witness the fact that they were not directly related to the events reported by the media in that period and that involved mainly the Syrian refugees.

From a content standpoint we tracked how hashtags usage is closely related to the relevant events. We have counted the frequency of each hashtag in each day, then performed a two-pass normalization. First we normalized the frequencies of hashtags on each day so as to avoid that days with lower recorded traffic are given less importance. Then we normalized each hashtag over the observed period, so that the values are comparable among different hashtags. We then measured the variance of the normalized frequencies, considering that hashtags with higher variance are those with a more unbalanced distribution among days. The hypothesis is that the unbalanced distribution is due to a close relation of the hashtag with a specific temporal event (usually one or two days).

Figure 12 shows the resulting twenty highest-variance hashtags.

The plot shows how this simple method allows us to quickly spot hot

topic in the observed stream of tweets and to correctly place them in time, i.e. the story of Alan Kurdi, drowned off the shores of Turkey the first days of September 2015.

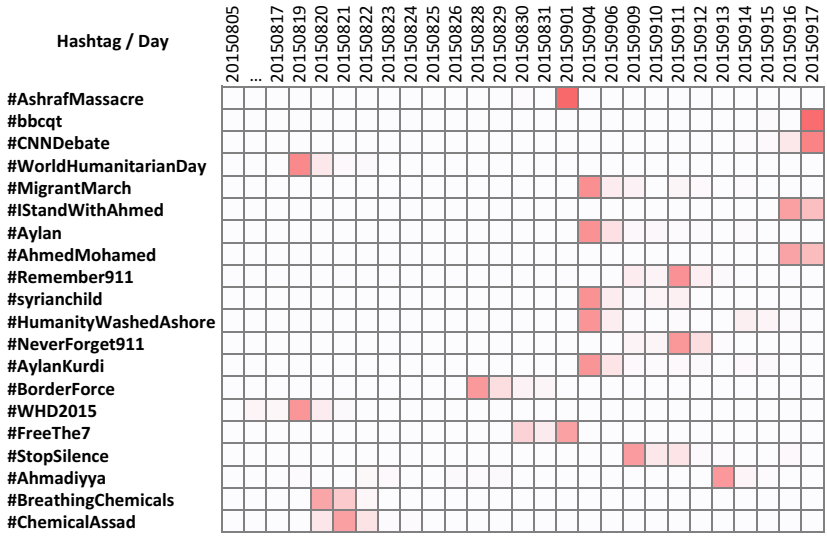


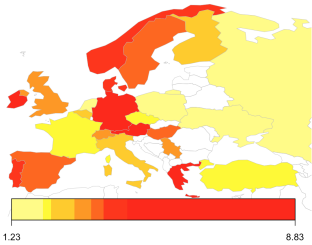
Figure 12: Highest-variance hashtags per day - intense red represents higher relative freq.

4.6.2 Sentiment analysis

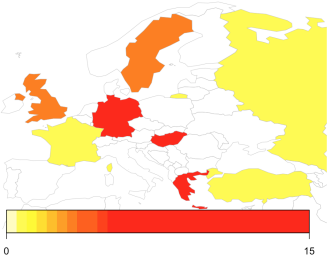
To answer the analytical question AQ2, we analyze the perception of the refugee crisis phenomenon by the European countries by exploiting the sentiment and location dimensions of the Twitter users in our dataset. To simplify the notation in the following we refer to $\mathcal{U}_L$ simply by $\mathcal{U}$.
Let us define ρ the ratio between the number of polarized users pro *refugees* and the number of users against *refugees*:

$$\rho = \frac{|\mathcal{U}_{c_+}|}{|\mathcal{U}_{c_-}|}$$

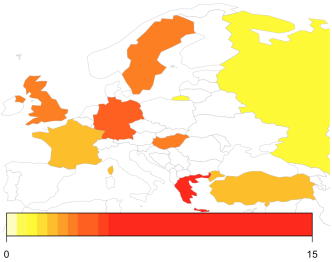
The index ρ gives a compact indication of the sentiment of a group of users.



(a) Global perception



(b) Internal perception



(c) External perception

Figure 13: Index ρ across European countries - red corresponds to a higher predominance of positive sentiment, yellow indicates lower ρ. (a) Refers to the whole dataset. (b) Is limited to users when mentioning locations in the their own country. (c) Is limited to users otherwise.

We first analyze the sentiment across the various countries, and we then differentiate between discussion about internal and external locations.

Sentiment by country

Figure 13(a) shows the value of ρ for users belonging to the different European countries.
We observe that Eastern countries in general are less positive than Western countries. In particular, Russia and Turkey have a low sentiment index probably because they are highly affected by the flow of arrivals.
On the contrary, countries like Germany and Austria are more positive and this can be confirmed by the news reporting their decision of opening borders to migrants.
Among western countries, France, UK, Italy and Netherlands have a low ρ index. In Italy the large amount of refugees arrived mainly through the sea directly from Lybia or Tunisia and the tone of the discussion is often characterized by negative notes. In France the sentiment confirms all worries about, probably, the situation of the "Calais jungle".
The situation in Greece appears very different. The sentiment is positive even though this country remains by far the largest single entry point for new sea arrivals in the Mediterranean, followed by Italy. Greece captured the attention of humanitarian organizations.
Countries like Ireland, Norway or Portugal are less interested by the phenomenon and therefore their perception might result more positive. Even for Spain ρ is not particularly low since the problem of refugees coming from Western Mediterranean was limited in number of people compared to central and eastern countries.

Internal and external country perception of the refugees crisis

In the following we study the perceived sentiment in relation to the user country. We denote as *internal perception* the sentiment of a user when mentioning his/her own country (or a city in his/her country). *External perception* refers to polarized tweets with no internal references.
Figure 13(b) shows the sentiment ratio ρ by country considering the

internal perception, thus tweets mentioning the country itself. The ρ computation refers to the users of a country who mentioned in their tweets the country itself (or indirectly a city in the country). We report countries for which we have a minimum amount of data. We can see that Russia, France and Turkey have a really low ρ index. We conjecture that the sentiment of a person, when the problem involves directly his/her own country, could be more negative since we are generally more critical when issues are closer to ourselves.
The external perception ratio is depicted in Figure 13(b). Comparing the two maps, we see that internal and external perception is stable for UK and Sweden. Other countries have a much lower internal sentiment ρ than external, and this is the case of France, Russia and Turkey. All these countries were indeed facing many critical problems due to the arrival of refugees to their borders. The case of *Calais* is one of the most significant examples which could explain the case of the low ratio in France. Germany, Hungary and Greece, on the contrary, have a better internal perception which might be due to the decision of Germany to open borders to allow many people to transit from Hungary to Germany, releasing the extremely difficult situation at the national borders.

Sentiment analysis: the UK case

In this section we focus on the sentiment analysis of UK citizens, as UK is the most represented country in our dataset and therefore a more detailed sentiment analyses can be done. Indeed, for UK we detail the results at the granularity of the city level.
Figure 14 (left) shows the number of polarized users $\mathcal{U}_L$ in the most represented cities of the country, with at least 100 polarized users in the dataset. We can see from the heatmap that there is a gradient of polarization from south to north in the sentiment. This could be due to the fact that the cities in the south were more involved in the welcoming process of refugees and this might have generated more discontent. On the other hand Scotland shows a more positive perception of the refugees migration.

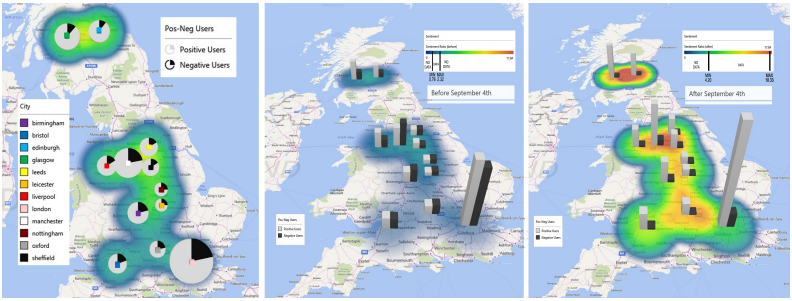


Figure 14: Positive and negative users for different cities in UK in all period (left) before (center) and after (right) September 4. In the infographic the pies/bars show the number of polarized positive and negative users by city and the heat map in background indicates the value of ρ for the cities considered in the legend. For some cities the tweets are not sufficient to compute polarization, therefore when the heat degrades to 0 it indicates no data.

From the time series of ρ for UK users we see an increase in the general sentiment ratio of the country after September 4. We find news[3] regarding that period from BBC and we think that the increase in the sentiment polarization could be be due mainly to the decision of the Prime Minister Cameron of acting with "head and heart" to help refugees. He allocated substantial amounts of money to humanitarian aid becoming, at that time, the second largest bilateral donor of aid to the Syrian conflict (after the US).

Figure 14 (middle) and (right), shows the comparison of the opinion in UK before and after September 4, respectively. We highlight again a gradient of polarization from north to south in both cases even though the sentiment ratio ρ before and after that day is completely overturned. After September 4 the spreading of positive news in UK increases the sentiment and the volume of relevant tweets in all the country and probably government position reflects the sentiment of a vast majority of users which show support to refugees in their digital statements.

[3]Sept, 04: `http://www.bbc.com/news/uk-34148913` – Sept, 16: `http://www.bbc.com/news/uk-34268604`

4.6.3 Mentioned location analysis

The last analysis we conduct aims at exploring AQ3 by studying the sentiment of the tweets when mentioning specific countries. We show how events impact differently the volume of tweets with positive or negative sentiment. Furthermore, we relate the sentiment changes to events.

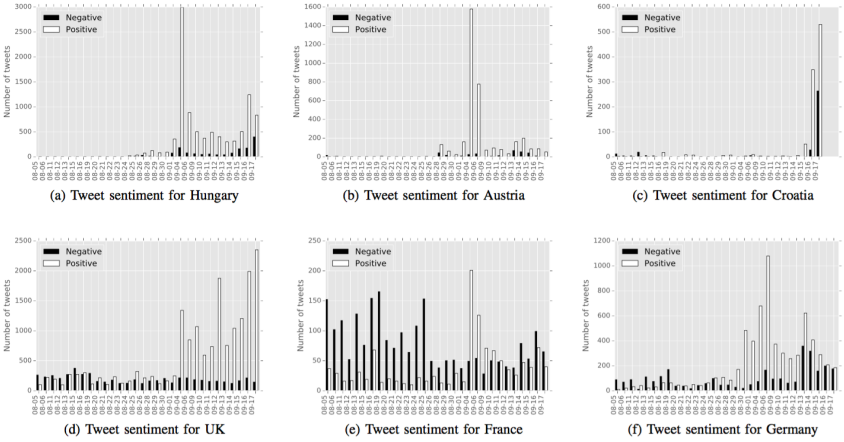


Figure 15: Tweet sentiment for country mentions per day.

Figure 15 (a-c) shows the sentiment of tweets when mentioning three of the countries most impacted by the refugees routes: Hungary, Austria and Croatia. We highlight an overall low number of mentions of these countries until the beginning of September.
In the case of Hungary and Austria there is a sudden increase in the beginning of September in the overall number of mentions, predominantly for c_+ with a relative increase in c_-. This is mostly due to the overall positive sentiment towards the events from the previous days (the Alan Kurdi story), but also due to positive news about migrants being allowed to cross the Austro-Hungarian border.
The negative sentiment appears, and continues to grow, until the middle

of September when c_- tends to increase more than c_+, due to tweets expressing negative feelings towards border controls in Austria (13-15 Sept) and Hungary arresting refugees crossing the border illegally (15-17 Sept). Croatia comes into play towards the end of our observation period when on the September 16th becomes a valid alternative to Hungary which closed its borders with Serbia. A similar analysis has been done for Greece, Macedonia and Serbia, but due to lack of space we are omitting here.
In Figure 15 (d-f) we look at the sentiment in relation to the mentions of UK, France and Germany.
Both UK and Germany are rather balanced between positive and negative tweets. Germany presents exceptions on certain days when a positive feeling arises in support to the sad incidents related to refugees. We notice that the official media news at the end of August reported that Germany was welcoming refugees, while UK started showing a positive sentiment after the dramatic facts of Alan Kurdi and the announcement of taking in 20,000 refugees by 2020.
France seems to have more negative feelings, probably due to the difficult situation in Calais and news about victims trying to across to UK, while a positive peak appears in correspondence to the Syrian boy news.

4.7 Conclusion

We proposed an multidimensional framework to analyze the spatial, temporal and sentiment aspects of a polarized topic discussed in an online social network. As a case study we used a Twitter dataset related to the Mediterrean refugee crisis. Besides enriching tweets with spatial and temporal information, one of the main contributions of this thesis is the sentiment enrichment methodology able to identify the polarity of users and tweets.
The combination of the sentiment aspects with the temporal and spatial dimension is an added value that allows us to infer interesting insights.
Our analysis revealed that European users are sensitive to major events and mostly express positive sentiments for the refugees, but in some cases

this attitude suddenly changes when countries are exposed more closely to the migration flow.

Chapter 5

Social influence and echo chambers

The results discussed in this chapter were published in (BCD+15; BCD+14b; BCD+14a).

- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., Quattrociocchi, W. (2015) Science vs conspiracy: collective narratives in the age of (mis) information. PLOS ONE

- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Quattrociocchi, W. (2014) Misinformation in the loop: the emergence of narratives in online social networks. In 13th Conference of the Italian chapter of AIS (Association for Information Systems). ITAIS 2014, November 21-22, Genova, Italy.

- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., Quattrociocchi, W. (2014) Sensing information-based communities in the age of misinformation. In European Conference on Complex Systems. ECCS 2014, September 22-26, Lucca, Italy.

5.1 Introduction

We have already pointed out in the previous chapters the relevance of opinions in the social interactions and how polarization clusters communities in sub-groups according to the belief system of the users.
Users, then, tend to discuss and access OSN content according to their *polarization class*, which means that in many contexts (politics, religion, science) they mainly consume information originated in the community they belong and they trust sources someway consistent with their belief system. In particular, in this chapter we want to discuss the debated hypothesis of *collective intelligence* that has been formulated in the past referring to democratic access and production of contents that Social Media enables compared to traditional media. According to our experiments people are influenced by their belief system and the consumption of content and the interactions follow often a self-contained dynamic, whose intensity is community dependent. We examine in this part of our dissertation the case of discussions about scientific news and conspiracy theories which are very frequent in Social Media to understand the dynamic of interactions and access to information from these sources.

5.2 Problem formulation

In this chapter we aim at answering the following research question:
Q4: *Do users interact with other users that do not share their belief system? How much isolated are communities in OSNs?*

In this chapter we elaborate on *echo chambers* in Social Media and how people tend to share information, ideas, or beliefs inside an *enclosed* system, where different or competing views are censored, disallowed, or otherwise underrepresented. We analyze the case of communities of supporters of science and conspiracy theories in Facebook, how they are structured and to what extend users interact out of their own community.

5.3 Contribution

In spite of the enthusiastic rhetoric about the so called collective intelligence unsubstantiated rumors and conspiracy theories - e.g., chemical trails, reptilians or the Illuminati - are pervasive in online social networks (OSN). In this chapter we study, on a sample of 1.2 million individuals, how information related to very distinct narratives, i.e., main stream scientific and conspiracy new - are consumed and shape communities on Facebook. Our results show that polarized communities emerge around distinct types of contents and usual consumers of conspiracy news result to be more focused and self-contained on their specific contents. To test potential biases induced by the continued exposure to unsubstantiated rumors on users' content selection, we conclude our analysis measuring how users respond to 4,709 troll information, i.e., parodistic and sarcastic imitation of conspiracy theories. We find that 77.92% of likes and 80.86% of comments are from users usually interacting with conspiracy stories, showing their higher attitude in consuming and trusting false information.

5.4 Related work

The World Wide Web has changed the dynamic of information transmission as well as the agenda-setting process (MS72). Facts, in particular when related to social relevant issues, mingle with half-truths and untruths to create informational blends (RM14; AP46). In such a scenario, as pointed out by (KQJ+00), individuals can be uninformed or misinformed and the role of corrections in the diffusion and formation of biased beliefs are not effective. In particular, in (BCDV+14) online debunking campaigns have been shown to create a reinforcement effect in usual consumers of conspiracy stories. In this work, we address user consumption patterns of information using very distinct type of contents, i.e., main stream scientific news and conspiracy news. The former diffuse scientific knowledge and the sources are easy to access. The latter aim at diffusing what is neglected by manipulated main stream media. Specifically, conspiracy theses tend to reduce the complexity

of reality by explaining significant social or political aspects as plots conceived by powerful individuals or organizations. Since these kinds of arguments can sometimes involve the rejection of science, alternative explanations are invoked to replace the scientific evidence. For instance, people who reject the link between HIV and AIDS generally believe that AIDS was created by the U.S. Government to control the African American population (SV09). The spread of misinformation in such a context might be particularly difficult to detect and correct because of the social reinforcement, i.e., people are more likely to trust the information someway consistent with their belief system (MM13; MR02; MS00; GW13; BSAS+12; Cen10; PEJ+09; QCL11; QPC09; BBEM11; QCS14). The growth of knowledge fostered by an interconnected world together with the unprecedented acceleration of scientific progress has exposed the society to an increasing level of complexity to explain reality and its phenomena. Indeed, a shift of paradigm in the production and consumption of contents has occurred, utterly increasing the volumes as well as the heterogeneity of available to users. Everyone on the Web can produce, access and diffuse contents actively participating in the creation, diffusion and reinforcement of different narratives. Such a large heterogeneity of information fostered the aggregation of people around common interests, worldviews and narratives.

Narratives grounded on conspiracy theories tend to reduce the complexity of reality and are able to contain the uncertainty they generate (Byf11; FCVH; HB11). They are able to create a climate of disengagement from mainstream society and from officially recommended practices (Bau97), e.g., vaccinations, diet, etc. Despite the enthusiastic rhetoric about the *collective intelligence* (Sur05; WBBP10) the role of socio-technical system in enforcing informed debates and their effects on the public opinion still remain unclear. However, the World Economic Forum listed massive digital misinformation as one of the main risks for modern society (How13).

A multitude of mechanisms animates the flow and acceptance of false rumors, which in turn create false beliefs that are rarely corrected once adopted by an individual (GW13; MR02; KGP00; AR98). The process

of acceptance of a claim (whether documented or not) may be altered by normative social influence or by the coherence with the belief system if the individual (ZCL+10; FNL11). A large body of literature addresses the study of social dynamics on socio-technical systems from social contagion up to social reinforcement (ORT10; UBMK12; LGK12; MBG+13; AG05a; Kle13; PEJ+09; QCL11; QPC09; BFJ+12; BHRG+11; Cen10; CFL09; QCS14; BnKVR03; FAcC; HMKW14; CAD+14).
Recently in (MRZ+14; BSZ+ar) it has been shown that online unsubstantiated rumors, such as the link between vaccines and autism, the global warming induced by chem-trails or the secret alien government, and main stream information, such as scientific news and updates, reverberate in a comparable way. The diffusion of unreliable contents might lead to mix up unsubstantiated stories with their satirical counterparts, e.g., the presence of sildenafil-citratum (the active ingredient of Viagra™) (sim14b) in chem-trails or the anti hypnotic effects of lemons (more than 45000 shares on Facebook) (sim14a; sim14c). In fact, there are very distinct groups, namely *trolls*, building Facebook pages as a caricatural version of conspiracy news. Their activities range from controversial comments and posting satirical contents mimicking conspiracy news sources, to the fabrication of purely fictitious statements, heavily unrealistic and sarcastic. Not rarely, these memes became viral and were used as evidence in online debates from political activists (Amb13).
In the work presented in this chapter we target consumption patterns of users with respect to very distinct types of information. Focusing on the Italian context and helped by pages very active in debunking unsubstantiated rumors, we build a list of scientific and conspiracy information sources on Facebook. Our dataset contains 271,296 post created by 73 Facebook pages. Pages are classified according to the kind of information disseminated and their self description in conspiracy news - alternative explanations of reality aiming at diffusing contents neglected by main stream information - and scientific news. For further details about the data collection and the dataset refer to the Methods section. Notice that it is not our intention claiming that conspiracy information are necessarily false. Our focus is on how communities formed around different information

and narratives interact and consume their preferred information.
In the analysis, we account for user interaction with respect to pages, public posts, i.e., likes, shares, and comments. Each of these actions has a particular meaning (ESL07; Joi08; VMCG09). A *like* stands for a positive feedback to the post; a *share* expresses the will to increase the visibility of a given information; and *comment* is the way in which online collective debates take form around the topic promoted by posts. Comments may contain negative or positive feedback with respect to the post. Our analysis starts with an outline of information consumption patterns and the community structure of pages according to their common users. We label polarized users - users which their like activity (positive feedback) is almost (95%) exclusively on the pages of one category - and find similar interaction patterns on the two communities with respect to preferred contents. According to literature on opinion dynamics (CFL09), in particular the one related to the Bounded confidence model (BCM) (DNAW01) - two individuals are able to influence each other only if the distance between their opinion is below a given distance - users consuming different and opposite information tend to aggregate into isolated clusters (*polarization*). Moreover, we measure their commenting activity on the opposite category finding that polarized users of conspiracy news are more focused on posts of their community and that they are more oriented on the diffusion of their contents, i.e., they are more prone to like and share posts from conspiracy pages. On the other hand, usual consumers of scientific news result to be less committed in the diffusion and more prone to comment on conspiracy pages. Finally, we test the response of polarized users to the exposure to 4709 satirical and demential version of conspiracy stories finding that, out of 3888 users labeled on likes and 3959 on comments, the most of them are usual consumers of conspiracy stories (80.86% of likes and 77.92% of comments). Our findings, coherently with (PJL13; EPDRH96; WK94) indicate that the relationship between beliefs in conspiracy theories and the need for cognitive closure, i.e., the attitude of conspiracists to avoid profound scrutiny of evidence to a given matter of fact, is the driving factor for the diffusion of false claims.

5.5 Data

5.5.1 Data collection

In this study we address the effect of the usual exposure to diverse verifiable contents on the diffusion of false rumors. We identified two main categories of pages: conspiracy news - i.e., pages promoting contents *neglected* by main stream media - and science news. We defined the space of our investigation with the help of Facebook groups active in debunking conspiracy theses. We categorized page according to their contents and their self description.
Concerning conspiracy news, their self description is often claiming the mission to inform people about topics neglected by main stream media. Pages like *Scienza di Confine* (eng: Frontier Science), *Lo Sai* (eng: You Know) or *CoscienzaSveglia* (eng: Awaken Consciousness) promote heterogeneous contents ranging from aliens, chemtrails, geocentrism, up to the causal relation between vaccinations and homosexuality. We do not focus on the truth value of their information but rather on the possibility to verify their claims. Conversely, science news, e.g *Scientificast*, *Italia unita per la scienza* (eng: Italy united for Science) are active in diffusing posts about the most recent scientific advances.

Table 15: Breakdown of Facebook dataset - The number of pages, posts, likes, comments, likers, and commenters for conspiracy and science news.

	Total	**Science News**	**Conspiracy News**
Pages	73	34	39
Posts	271,296	62,705	208,591
Likes	9,164,781	2,505,399	6,659,382
Comments	1,017,509	180,918	836,591
Likers	1,196,404	332,357	864,047
Commenters	279,972	53,438	226,534

To our knowledge, the final dataset is the complete set of all scientific and

conspiracist information sources active in the Italian Facebook scenario. In addition, we identify two pages posting satirical news with the aim of mocking usual rumors circulating on line by adding satirical contents.
The entire data collection process has been carried out exclusively through the Facebook Graph API (Fac13). The pages from which we downloaded data are public Facebook entities (can be accessed by virtually any user). The resulting dataset is composed of 73 public pages divided in scientific and conspiracist news for which we downloaded all the posts (and their respective users interactions) over a timespan of 4 years (2010 to 2014).
The breakdown of the data is presented in Table 15. The first category includes all pages diffusing conspiracy information: pages which disseminate controversial information, most often lacking supporting evidence and sometimes contradictory of the official news (i.e., conspiracy theories). The second category is that of scientific dissemination including scientific institutions and scientific press having the main mission to diffuse scientific knowledge.

5.5.2 List of pages

The full list of pages related to scientific news and conspiracy theories can be found in (BCD+15).

5.6 Method and results

We start our analysis by characterizing users' interaction patterns with respect to different kind of contents. Then, we label typical users according to the kind of information they are usually exposed to and validate their tolerance with respect to information that we know to be false as they are a parodistic imitation of conspiracy stories containing fictitious and heavily unrealistic statements.

5.6.1 Preliminaries and definitions

Statistical Tools. To characterize random variables, a main tool is the probability distribution function (PDF), which gives the probability that a

random variable X assumes a value in the interval $[a, b]$, i.e., $P(a \leq X \leq b) = \int_a^b f(x)dx$. The cumulative distribution function (CDF) is another important tool giving the probability that a random variable X is less than or equal to a given value x, i.e., $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$. In social sciences, an often occuring probability distribution function is the Pareto's law $f(x) \sim x^{-\gamma}$, that is characterized by power law tails, i.e., by the occurrence of rare but relevant events. In fact, while $f(x) \to 0$ for $x \to \infty$ (i.e., high values of a random variable X are rare), the total probability of rare events is given by $C(x) = P(X > x) = \int_x^{\infty} f(y)dy$, where x is a sufficiently large value. Notice that $C(x)$ is the Complement to the CDF (CCDF), where complement indicates that $C(x) = 1 - F(x)$. In order to better visualize the behavior of empirical heavy-tailed distributions, we recur to log-log plots of the CCDF.

Bipartite Networks and Community Detection. We consider a bipartite network having as nodes users and affiliation the Facebook pages. A comment to a given information posted by a page determines a link between a user and a page. More formally, a bipartite graph is a triple $\mathcal{G} = (A, B, E)$ where $A = \{a_i \,|\, i = 1 \dots n_A\}$ and $B = \{b_j \,|\, j = 1 \dots n_B\}$ are two disjoint sets of vertices, and $E \subseteq A \times B$ is the set of edges. Edges exist only between vertices of the two different sets A and B. The bipartite graph $\mathcal{G}$ is described by the matrix M defined as

$$M_{ij} = \begin{cases} 1 & if\ an\ edge\ exists\ between\ a_i\ and\ b_j \\ 0 & otherwise \end{cases}$$

For our analysis we use the co-occurrence matrices $C^A = MM^T$ and $C^B = M^T M$ that count, respectively, the number of common neighbors between two vertices of A or B. C^A is the weighted adjacency matrix of the co-occurrence graph $\mathcal{C}^A$ with vertices on A. Each non-zero element of C^A corresponds to an edge among vertices a_i and a_j with weight P^A_{ij}.
To test the community partitioning we use two well known community detection algorithms based on modularity (BGLL08a; CNM04). The former algorithm is based on multi-level modularity optimization. Initially, each vertex is assigned to a community on its own. In every step,

vertices are re-assigned to communities in a local, greedy way. Nodes are moved to the community in which they achieve the highest modularity. Differently, the latter algorithm looks for the maximum modularity score by considering all possible community structures in the network. We apply both algorithms to the bipartite projection on pages.

Labeling algorithm. The labeling algorithm can be described as thresholding strategy on the total number of users likes. Considering the total number of likes of a user L_u on both posts P in categories S and C. Let l_s and l_c define the number of likes of a user u on P_s or P_c, respectively denoting posts from scientific and conspiracy pages. Then, we will have the total like activity of users on one category expressed as $\frac{l_s}{L_u}$. Fixing a threshold θ we can discriminate users with enough activity on one category. More precisely, the condition for a user to be labeled as a polarized user in one category x is $\frac{l_x}{L_u} > \theta$.
In Figure 16 we show the number of polarized users as a function of θ. Both curves decrease with a comparable rate.

5.6.2 Consumption patterns on science and conspiracy news

Our analysis starts by looking at how Facebook users interact with contents from pages of conspiracy and mainstream scientific news.
Figure 17 shows the empirical complementary cumulative distribution function (CCDF) for likes (intended as positive feedbacks to the post), comments (a measure of the activity of online collective debates), and shares (intended as the will to increase the visibility of a given information) for all posts produced by the different categories of pages. Distributions of likes, comments, and shares on both categories are heavy-tailed.
A post sets the attention on a given topic, then a discussion may evolve in the form of comments. To further investigate users consumption patterns, we zoom in at the level of comments. Such a measure is a good approximation of users attention with respect to the information reported on by the post.
Figure 18 shows CCDF of the posts lifetime, i.e., the temporal distance be-

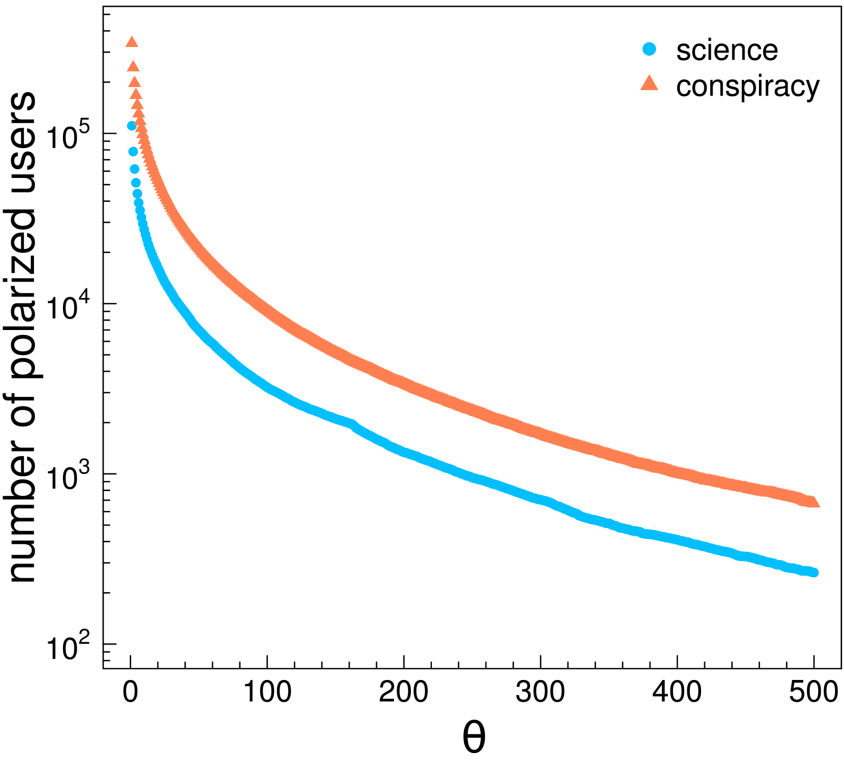


Figure 16: Polarized users and activity - Number of polarized users as a function of the thresholding value θ on the two categories.

tween the first and the last comment for each post from the two categories of pages. Very distinct kinds of contents have have a comparable lifetime. To account for the distinctive features of the consumption patterns related to different contents, we focus on the correlation of combination of users' interactions with posts. Likes and comments have a different meaning from a user viewpoint. Notice that, cases in which they are motivated by ironic reasons are impossible to detect. In order to compute the correlation among different actions, we use the Pearson coefficient, i.e., the covariance of two variables (in this case couples of action) divided by the product of

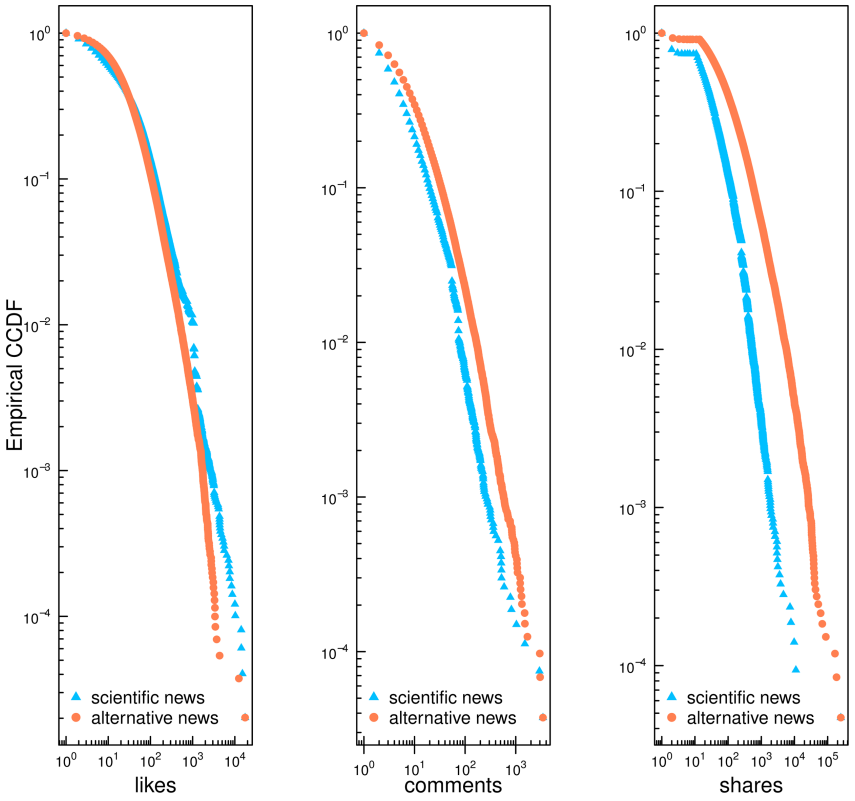


Figure 17: Users activity - Empirical complementary cumulative distribution function (CCDF) of users activity (like, comment and share) for post grouped by page category. The distributions are indicating heavytailed consumption patterns for the various pages.

their standard deviations. In Table 16 we show the Pearson correlation for user couple of actions on posts (likes, comments and shares). The values are the correlations between all the possible combinations among the three variables: number of likes, comments, shares for each post in each pages of different categories (science and conspiracy pages).

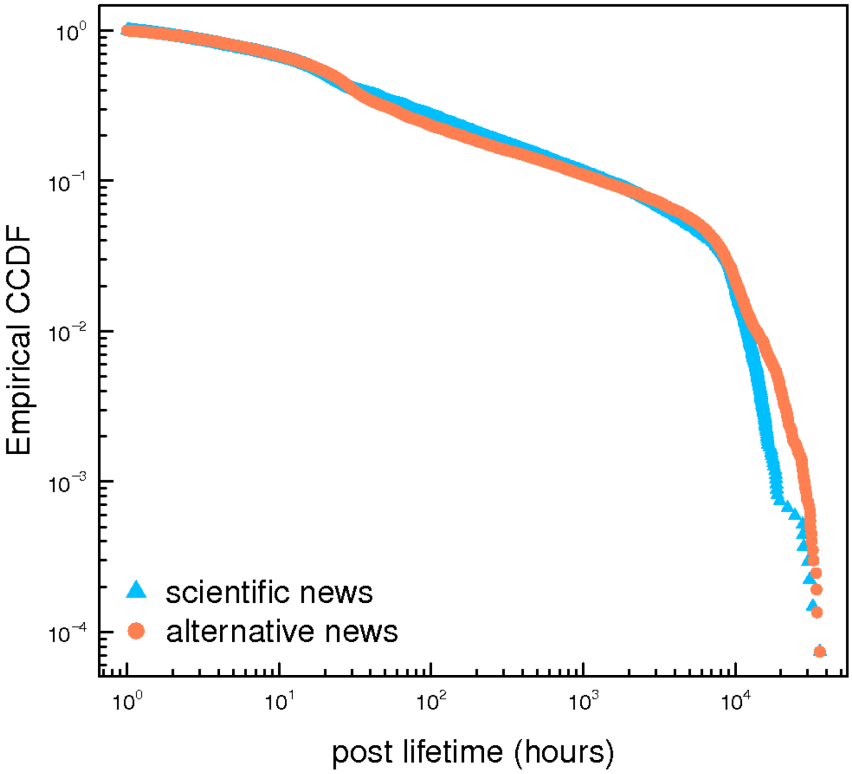


Figure 18: Post lifetime - Empirical complementary cumulative distribution function (CCDF), grouped by page category, of the temporal distance between the first and last comment to each post. The life time of posts in both categories is similar.

Correlation values for posts of conspiracy news have higher values than those in science news, indicating a preference of conspiracy users to promote their liked contents in many ways. This finding is consistent with (PJL13; EPDRH96; WK94) which state that conspiracists need for cognitive closure, i.e., they are more likely to interact with conspiracy based theories and have a lower trust in other information sources. Qualitatively different information are consumed in a comparable way.

Table 16: Users actions - Correlation (Pearson coefficient) between couple of actions to each post in scientific and conspiracy news. Posts from conspiracy pages are more likely to be liked and shared by users, indicating a major commitment in the diffusion.

	Likes/Comments	**Likes/Shares**	**Comments/Shares**
Science	0.523	0.218	0.522
Conspiracy	0.639	**0.816**	0.658

However, zooming in at the combination of actions we find that users of conspiracy pages are more prone to share and like on a post. Such a latter result indicates a higher level of commitment of consumers of conspiracy news. They are more oriented to the diffusion of conspiracy related topics that are - according to their belief system - neglected by main stream media and scientific news and consequently very difficult to verify. Such pattern, oriented to diffusion of conspiracy news, opens to interesting about the pervasiveness of unsubstantiated rumors in online social media.

5.6.3 Information-based communities

The classification of pages in science and conspiracy related contents is grounded on their self-description and on the kind of promoted content. We want to understand if users engagement across very distinct contents shapes different communities around contents. We apply a network based approach aimed at measuring distinctive connectivity patterns of these information-based communities, i.e., users consuming information belonging to the same narrative. In particular, we transform data in order to have a bipartite network of pages and users. We consider now the projection of the bipartite graph on the pages, i.e., two pages are connected if a user liked a post from both of them.
In Figure 19 we show different clustering of pages (orange for conspiracy and azure for science). In the first panel, memberships are given according to our categorization of pages (for further details refer to Section 5.5). The second panel shows the page network with membership given by

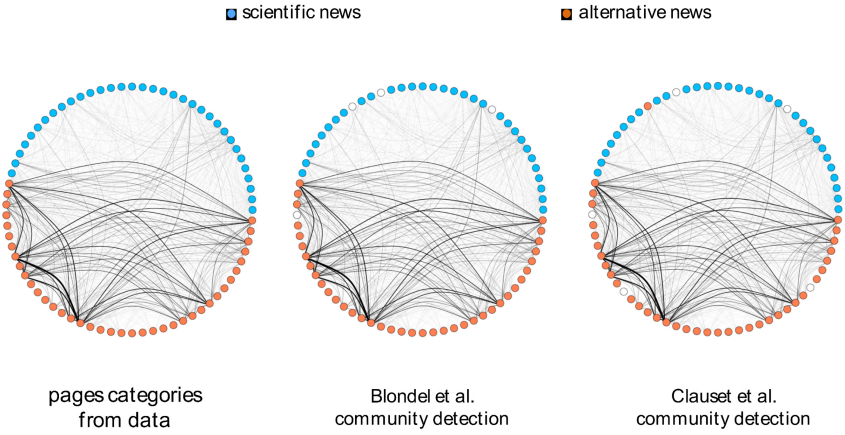


Figure 19: Page network - The membership of 73 pages as a) identified by means of their self-description, b) by applying the multi-level modularity optimization algorithm, and c) by looking at the maximum modularity score. Community detection algorithms based on modularity are good discriminants for community partitioning.

applying the multi-level modularity optimization algorithm (BGLL08a). In the third panel, membership is obtained by applying an algorithm that looks for the maximum modularity score (CNM04).
These findings indicates that connectivity patterns, in particular the modularity, between the two categories of pages differ. Since we are considering users' likes on the pages' posts, this aspect is pointing out a higher mobility of users of across pages of the conspiracy category.

5.6.4 Polarized users and their interaction patterns

In this section we focus on the users engagement across the different contents. Hence, we label users by means of a simple thresholding algorithm accounting for the percentage of likes on one or the other category. Notice that the choice of the *like* as a discriminant is grounded on the fact that generally such an action stands for a positive feedback to a post (VMCG09).

We consider a user to be polarized in a community when the number of his/her likes with respect to his/her total like activity on one category - scientific or conspiracy news - is higher than 95% (for further details about the algorithm refer to Section 5.5). We identify 255,225 polarized users of scientific pages, resulting to be the 76,79% of users interacted on scientific pages, and 790,899 conspiracy polarized users, the 91,53% of users interacting with conspiracy pages in terms of liking. Users activity across pages is highly polarized.
According to literature on opinion dynamics (CFL09) in particular the one related to the Bounded Confidence Model (BCM) (DNAW01) - two nodes are able to influence each other only if the distance between their opinions is below a given distance - users consuming different and opposite information tend to form polarized clusters. The same hold if we look at commenting activity of polarized users inside and outside their community. In particular, those users that are polarized on conspiracy news tend to interact especially in their community both in terms of comments (99%) and likes. Users polarized in science tend to comment slightly more outside their community. Results are summarized in Table 17.

Table 17: Activity of polarized users (S = Science, C = Conspiracy) - Quantity and percentage of classified users for each category and their commenting activity on the category in which they are classified and on the opposite category. Users polarized on conspiracy pages tend to interact especially in their community both in terms of comments and likes. Users polarized in science are more active elsewhere.

	Users labeled	**% Users labeled**	**% Comments (own cat.)**	**% Comments (other cat.)**	**Comments (both)**
S	$255,225$	76.79	90%	10%	$140,057$
C	$790,899$	91.53	99%	1%	$648,183$

Figure 20 shows the CCDF for likes and comments of polarized users. Despite the very profound different nature of contents, consumption patterns are nearly the same both in terms of likes and comments. This finding

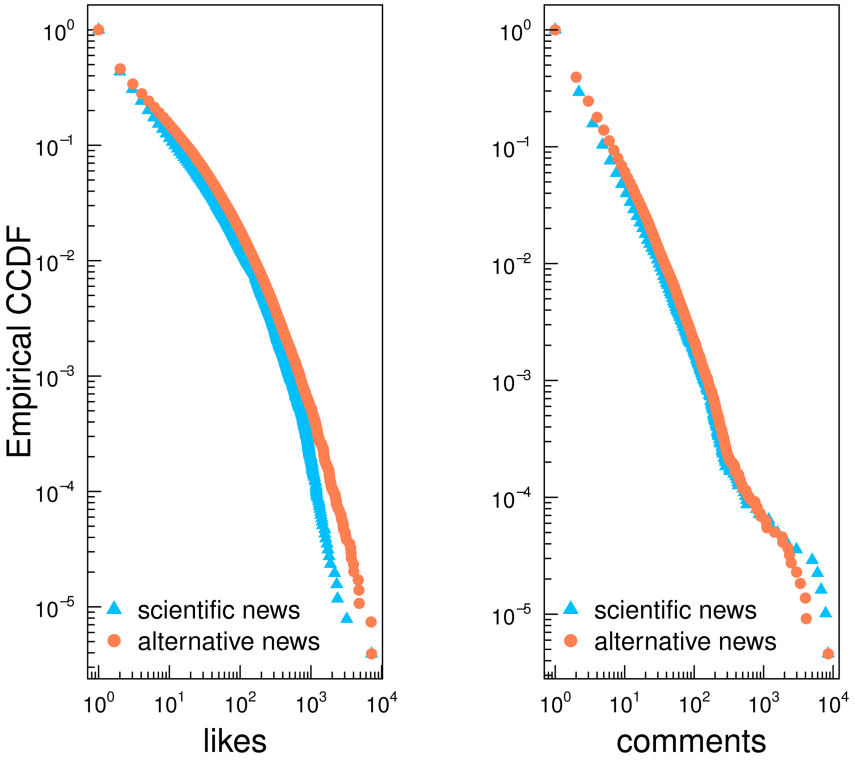


Figure 20: Consumption patterns of polarized users - Empirical complementary cumulative distribution function (CCDF) for likes and comments of polarized users.

suggests that very engaged users of different and clustered communities formed around different kind of narratives consume their preferred information in a similar way.

As a further investigation, we focus on the post where polarized users of both communities commented. Hence, we select the set of posts on which at least a polarized user of each of the two communities has commented. We find polarized users of communities debating on $7,751$ posts ($1,991$ from science news and $5,760$ from conspiracy news). The

post at the interface, where the two communities discuss are mainly on the conspiracy side. As shown in Figure 21, polarized users of scientific news

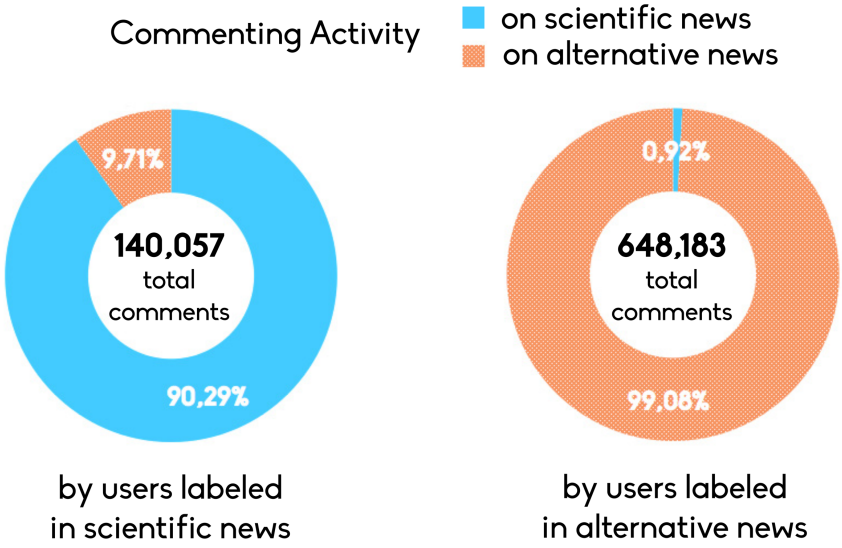


Figure 21: Activity and communities - Posts on which at least a member of each the two communities has commented. The number of posts is 7,751 (1,991 from scientific news and 5,760 from conspiracy news). Here we show the commenting activity in terms of polarized users on the two categories.

made 13,603 comments on post published by conspiracy news (9.71% of their total commenting activity), whereas polarized users of conspiracy news commented on scientific posts only 5,954 times (0.92% of their total commenting activity, i.e., roughly ten times less than polarized users of scientific news).

5.6.5 Response to false information

On online social networks, users discover and share information with their friends and through *cascades* of reshares information might reach a large number of individuals. Interesting is the popular case of Senator Cirenga's (Cir14b; Cir14a) law proposing to fund policy makers with 134

billion of euros (10% of the Italian GDP) in case of defeat in the political competition. This was an intentional joke with an explicit mention to its satirical nature. The case of Senator Cirenga became popular within online political activists and used as an argumentation in political debates (Amb13).
Our analysis showed that users tend to aggregate around preferred contents shaping well defined groups having similar information consumption patterns. Our hypothesis is that the exposure to unsubstantiated claims (that are pervasive in online social media) might affect user selection criteria by increasing the attitude to interact with false information. Therefore, in this section we want to test how polarized users usually exposed to distinct narrative - - one that can be verified (science news) and one that by definition is almost impossible to check - - interact with posts that are deliberately false.
To do this we collected a set of troll posts - i.e., paradoxical imitations of conspiracy information sources. These posts are clearly unsubstantiated claims, like the undisclosed news that infinite energy has been finally discovered, or that a new lamp made of actinides (e.g., plutonium and uranium) might solve problems of energy gathering with less impact on the environment, or that the chemical analysis revealed that chem-trails contains sildenafil citratum (the active ingredient of Viagra^{TM}).
Figure 22 shows how polarized users of both categories interact with troll posts in terms of comments and likes. We find that polarized users of conspiracy pages are more active in liking and commenting on intentionally false claims.

5.7 Conclusion

Recently in (MRZ+14; BSZ+ar) has been shown that unsubstantiated claims reverberate for a timespan comparable to the one of more verified information and that usual consumers of conspiracy theories are more prone to interact with them. Conspiracy theories find on the internet a natural medium for their diffusion and, not rarely, trigger collective counter-conspirational actions (AG12; LCOM13). Narratives grounded

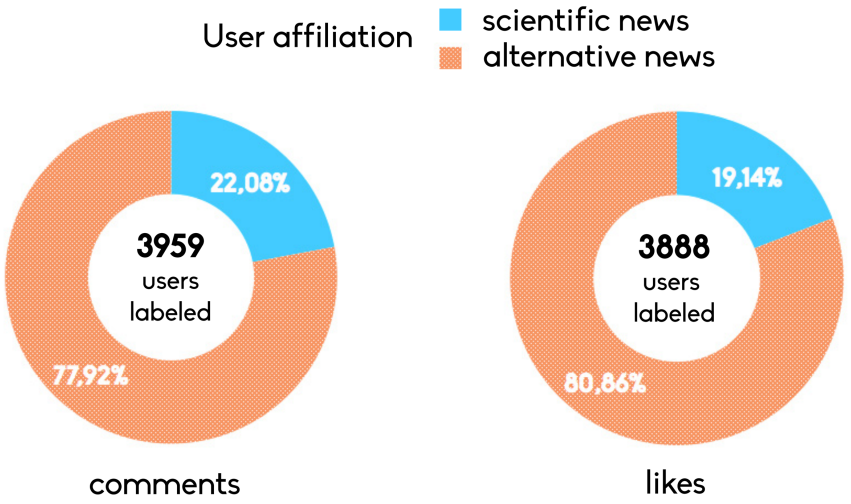


Figure 22: Polarized users on false information - Percentage of comments and likes on intentional false memes posted by a satirical page from polarized users of the two categories.

on conspiracy theories tend to reduce the complexity of reality and are able to contain the uncertainty they generate (Byf11; FCVH; HB11).
We studied how users interact with information related to different (opposite) narratives on Facebook. Through a thresholding algorithm we label polarized users on the two categories of pages identifying well shaped communities. In particular, we measure commenting activity of polarized users on the opposite category, finding that polarized users of conspiracy news are more focused on posts of their community and their attention is more oriented to diffuse conspiracy contents. On the other hand, polarized users of scientific news are less committed in the diffusion and more prone to comment on conspiracy pages. A possible explanation for such a behavior is that the former want to diffuse what is neglected by main stream thinking, whereas the latter aims at inhibiting the diffusion of conspiracy news and proliferation of narratives based on unsubstantiated claims. Finally, we test how polarized users of both

categories responded to the inoculation of $4,709$ false claims produced by a parodistic page, finding polarized users of conspiracy pages to be the most active.

These results shown confirm the formation of communities in OSNs according to the opinion polarization and the presence of *echo chambers*. In particular, we found that the cognitive closure is present in different communities (scientific groups and conspiracy theories believers), with a different degree: for instance, conspiracists are more self-contained and more prone to react diffusing fake injected information which is consistent with their belief system.

Chapter 6

Controversy: detection and analysis

The results discussed in this chapter were submitted to (CGGL17).

- Coletto, M., Garimella, K., Lucchese, C., Gionis, A. (2017). A motif-based approach for identifying controversy. Submitted to the 11th International Conference on Web and Social Media. ICWSM 2017, May 15-18, Montreal, Canada.

6.1 Introduction

Even though many studies have been devoted to understand different aspects of the social network structure and its function, such as, community structure (For10), information spreading (BRMA12), information seeking (KLPM10), link prediction (LNK07), etc., much less work is available on analyzing online discussions and public debates. In the previous chapter we have pointed out how users tend to interact mostly with other users who share their belief system. This tendency creates *echo chambers* which may result in weaken the idea of Social Media as a space for democratic collective intelligence: the access to the content follows user polarization and in most of the cases it reinforces his/her believes.

For some highly controversial topics (e.g., politics, religion, ethics) even though users prefer to get informed though polarized content originated in the communities they belong to, they like to share their affiliations, believes, ideals, convictions with external users persuading them in joining their belief system or supporting, criticizing an event, a group, a party or a specific person.
In the case of scientific communities in Facebook, for instance, we have already detected the attempt their members to interact with content that belong to other communities (i.e., conspiracists) in order to debunk it, even though the consequence is often the reinforcement in usual consumers of conspiracy stories (BCDV$^+$14).
The concept of *controversy* then is connected with the concept of *polarization* but it describes the interaction among two or more opponent polarized communities that discuss together, often with heated tones. Highly *polarization* does not always imply controversy because in some situations the different polarized communities are apart from each other and they do not interact at all, reinforcing the *echo chamber* effect. On the other hand, high controversy and quantity of interactions among members of different polarized communities are not a proof of absence of *echo chambers* because the final effect is often the reinforcement of user own idea instead of a change of opinion due to the human cognitive closure.

6.2 Problem formulation

In this chapter we aim at answering the following research question:

Q5: *How do users of different polarized communities interact? What is controversy and how can me measure it? Can we automatically detect controversy without looking at the content delivered among users?*

People like to express their opinions in favor or against a particular idea, supporting or criticizing a particular political candidate or a party. In these cases the polarized communities interact each other creating controversy. In this chapter we explore the concept of *controversy* and, by presenting the state of the art in detecting methods, we discuss a novel approach

which quantifies controversy without any information about the content of the conversation. The features used in the proposed machine-learning model take into consideration the social network, time-based actions, and - most importantly - conversational interaction patterns.

6.3 Contribution

Identifying controversial topics is useful for exploring the space of public discourse and understanding the issues of current interest. Thus, a number of recent studies have focused on the problem of identifying controversy in social media. Most of these studies are based on the analysis of textual content or rely on global network structure. Such approaches have strong limitations due to the difficulty of understanding natural language, especially in short texts, or due to the difficulty of extracting global network structure and developing appropriate measures.
In this chapter we study the problem of identifying controversies in social media, one of the many different aspects of analyzing online discussions and understanding how people participate in those. As a motivating application for the problem we address, consider a tool that identifies controversial topics and suggests to users the most relevant, according to their interests, so as users can browse them and participate in some of the discussions, if they wish. The underlying assumption here is that controversial topics can be interesting and intellectually stimulating as there are trade-offs to consider and opposing points of view.
The problem of studying controversy in social media has recently drawn some attention (GDFMGM16; LCN15). However, as this is a difficult problem, involving processing of human language and network dynamics, existing studies have limitations. For example, many papers study controversy in very controlled case studies, or focus on a predefined topic, most typically politics (CRF$^+$11b), for which they employ auxiliary domain-specific sources and datasets. In other cases, proposed approaches are based on content-based analysis (MZDC14), which has several limitations, as well, due to the ambiguity of the language and the fact that models become language-dependent and topic-dependent.

Instead, in this chapter we aim to identify controversies on *any* topic, discussed in *any* language. Given this objective, our approach is based on the analysis of the *network structure*. In this sense, this chapter is related to the recent work of Garimella et al. (GDFMGM16), who also aim at identifying controversies in the wild, independent of topic or language. In that work, the authors focus on a topic defined by a single hashtag, and then analyze the retweet network after partitioning it into two clusters (the two sides of controversy). The obvious limitation is that they assume that a topic partitions the network into two clusters (while none, or more than two clusters, may be present), and that it is computationally feasible to identify those clusters. In the work presented in the current chapter, we overcome those limitations by analyzing local network patterns (*motifs*), and thus, making no assumption about the global cluster structure of the network, neither about our ability to detect network clusters.

Moreover, note that the separation of the retweet network in communities does not always reflect controversy; it may also mean that a hashtag is used in two communities with different acceptations. Our model catches antagonism in the conversation and, in fact, we find that some hashtags (#germanwings, #onedirection) that were detected as not controversial by previous studies, contain controversial discussions.

Finally, in the work of Garimella et al. (GDFMGM16) the approach of detecting controversy is static and is based on analyzing the retweets of a given hashtag. In our case we focus on the analysis of the discussions generated by those tweets. This allows us to discover potentially controversial sub-topics that may be present within an otherwise non-controversial topic.

The novel contributions of this chapter are as follows.

We propose the use of motifs extracted from the user reply and friendships graphs to detect controversial threads of discussion in online social networks. The proposed motifs can be easily computed as they encompass interactions among two or three users only. Being graph-based, such motifs are language independent and topic independent: they can be applied to investigate interactions in social networks without any additional domain knowledge.

We measure the prediction power of the proposed motifs on a collection of Twitter data. We found that local motifs can improve the accuracy of frequently used graph-based features (e.g., cascade depth, inter-reply time) achieving an accuracy of 85%. We claim that such motifs are able to model both user homophily, through the friendship graph, and user interest in discussing specific topics even beyond their social circles, through the reply graph.
Finally, the proposed motifs, being local to two or three users, allow a fine-grained analysis of the evolution of a discussion over time and of the interactions among its users. In fact, we found that non controversial conversations happen to become controversial either limitedly to a subtree of the discussion thread, or globally due for instance to external events such as news.
In this chapter we show that it is possible to detect controversy in social media by exploiting network motifs, i.e., local patterns of user interaction. The proposed approach allows for a language-independent and fine-grained analysis of user discussions and their evolution over time. Network motifs can be easily extracted both from user interactions and from the underlying social network, and motif-based measures for controversy identification are conceptually simple to define and very efficient to compute.
We assess the predictive power of motifs on a manually labeled twitter dataset. In fact, a supervised model exploiting motif patterns can achieve 85% accuracy, with an improvement of 7% compared to structural, propagation-based and temporal network features. Finally, thanks to the locality of motif patterns, we show that it is possible to monitor the evolution of controversy in a conversation over time thus discovering changes in user opinion.

6.4 Related work

Controversy and polarization The analysis of controversy on the web and social media has received considerable attention in recent years, with a number of papers studying controversy on general web pages (DHA13),

blogs (AG05b), online news (CJM10; MZDC14), and social media (AQC14; GDFMGM16).

The existence of polarization on social media was first studied by Adamic et al. (AG05b) who identified a clear separation in the hyperlink structure of political blogs. Conover et al. (CRF$^+$11b) studied this phenomenon on Twitter, evaluating the polarization on the retweet network. In a more recent work, Garimella et al. (GDFMGM16) showed that the polarized structure in the retweet graph extends beyond politics. They also proposed algorithmic methods to measure the amount of controversy on a topic, by considering the structure of the network formed by retweets and followers. In a similar spirit, Guerra et al. (GMJCK13) considered a measure based on boundary connectivity patterns in order to identify if a discussion is controversial. Other approaches have also been proposed to identify controversy on social media at a *user* level. For example, BiasWatch is a weakly-supervised approach fusing content and network data to infer user polarity (LCN15).

Controversies are inherently dynamic. Non-controversial topics could become controversial and vice-versa. Morales et al. (MBLB15) present an approach based on label propagation in order to quantify the level of controversy in the network at a certain time instance. They apply their measure on Twitter data from Venezuela over a long period and showed that they can capture real-life shifts in polarization. Coletto et al. (CLOP16) proposed an approach for jointly tracking user polarity and topic evolution. The method proposed in this chapter can handle the dynamic nature of a controversial topic, as seen in Section 6.7.2.

Graphs (reply graphs) are used to represent the dynamic nature of information and discussion threads in a network. Various studies have proposed methods to extract and analyze conversation graphs on Twitter (CAB$^+$12; NTO$^+$16). Those studies analyze various types of conversation graphs, such as *long path-like reply trees*, *large star-like trees*, and *long irregular trees*. They also show that paths are making up to 60% of the reply graphs. In our experiments, we observe that reply graphs of Twitter discussions are composed by a majority of star-like trees. For

controversial discussions, we additionally detect long trees with multiple branches indicating the different threads of the discussions — e.g., see Figure 23.
Analysis of conversation graphs in rumor and misinformation spreading has shown that information flow in the network gives rise to certain types of local patterns (DDLMM13; CMP11). However, to our knowledge, this is the first attempt to study the role of network motifs in the context of identifying controversy in social media.

Motifs indicate patterns of interactions/interconnections in complex networks. The work of Milo et al. (MSOI$^+$02) was one of the first to analyze the occurrence of different motifs in networks arising in a wide range of fields, from biochemistry to engineering. Their finding that "*motifs may thus define universal classes of networks*" is one of our motivations for exploring simple interaction patterns related to controversy.
In the context of social networks, motifs may indicate a specific function or role of certain nodes. For example, network motifs have been used recently to explain higher-order network organization, and subsequently, use this information to cluster networks (BGL16).

Conversation textual analysis The problem of detecting disagreement in conversation text was recently studied by Allen et al. (ACN14), who use rhetorical structure features to identify disagreement. They claim that this is a difficult task, even for humans.
Most related to our work is the paper by Chen et al. (CB13), who study when, why, and how a conversation is initiated by a controversy. Their main hypothesis is that a controversy generally brings up interest and discomfort in users, and when the former is higher, a controversy causes a conversation, while otherwise, the likelihood of starting a conversation is smaller. Supporting evidence for this hypothesis is obtained by analyzing an online news website.
Furthermore, language-analysis tools have been used widely to determine the emotional tone of a conversation (KPV$^+$14), e.g., whether a message is partial/impartial (ZGDNM16), subjective/ objective, positive/negative

(BF10), etc.
All the different methods discussed above use only textual information. Even though the use of text features is orthogonal to our method, and they can be added separately, we chose not to do so explicitly, since text-analysis tools are language dependent, and since we are mainly interested in contrasting network motifs with other network-structure features. Nevertheless, as said, text analysis can be incorporated easily in our framework, and we leave this direction for future work.

In summary, the proposed method makes use of motifs in a combined user graph (follow relation) and reply graph (conversation) to identify controversy in a domain- and language-agnostic way. To our knowledge, this is the first study of this kind.

6.5 Data

Our main source of data is a carefully-curated set of popular Twitter pages for which we can easily identify a ground truth (if the page is controversial or not). The list is chosen to cover a wide range of domains (news, politics, celebrity, gossip, entertainment) and languages. Contrary to most previous papers on controversy, our study is not restricted to the political domain. The way we choose our seeds and collect the data is generic enough, and can be emulated on other social networks, such as Facebook.

Dataset1 - Twitter pages For each page, we gather a list of the 200 most recent tweets (in May 2016) and manually evaluate a sample of the tweets to check if they are controversial or not. We consider only the accounts whose tweets are mainly controversial or mainly not controversial, and we discard all the tweets from accounts with less than 90% controversial/non-controversial tweets. The final list of the 12 controversial and 7 non-controversial selected pages are shown in Table 18. In the subsequent analysis, we use the page as a label for all the tweets in that page, i.e., a tweet is deemed controversial (non-controversial) if it originates from a controversial (non-controversial) page.

Table 18: List of Twitter pages used in our study (Dataset1)

Controversial	@tedcruz, @mov5stelle, @brexitwatch @barackobama, @realdonaldtrump @wikileaks, @berniesanders, @cnnbrk @bbcworld, @hillaryclinton, @potus
Non Controversial	@coldplay, @justinbieber, @cristiano @adele, @chanel, @xbox, @nba

For each tweet in each of these pages, we reconstructed the generated discussion thread by recursively crawling the tweet's replies. We select tweets that generate a conversation involving more than k users (including the author of the original post). The reply tweets are often in a different language than the language of the original tweet, including Arabic, Russian, and others. Table 19 reports the number of posts we collect with the above procedure, with $k = 2, 3, 10$. The table reports the average number of users who take part in the conversation, and the average number of reply links among the users. Each collected tweet generates a network of replies that involves on average about 100 users.

Dataset2 - Twitter hashtags In order to be consistent with the recent literature, we also collect tweets based on controversial and non-controversial hashtags, in particular the ones used by Garimella et al. (GDFMGM16). We use four controversial (#beefban, #baltimore, #netanyahuspeech and #russia_march) and four non-controversial hashtags (#germanwings, #onedirection, #sxsw, #ultralive). For each hashtag we collect 200 tweets. For each tweet we collect all the replies and build the dataset in the same way that was described before. Statistics on this dataset are reported in Table 19.

We note that, upon manual inspection, for many hashtags in the abovementioned dataset, there is a mix of different behaviors depending on the context in which the hashtag is used in the tweets. Some are predominantly controversial or non-controversial, while others are mixed. Dataset2 is used as an additional test set for our model trained on Dataset1 to assess the controversial nature of popular hashtags (Section 6.7.3).

Table 19: Data statistics

Dataset1: ***Twitter pages***

Filtering	Posts	Avg. Users	Avg. Replies
> 2 users	1202	108	118
> 3 users	1175 (97%)	110	120
> 10 users	1046 (87%)	123	134

Dataset2: ***Twitter hashtags***

Filtering	Posts	Avg. Users	Avg. Replies
> 2 users	1302	32	34
> 3 users	1211 (93%)	34	36
> 10 users	699 (54%)	54	57

6.6 Method and model

Given a social network we are interested in modeling the interactions among users and the dynamics incurring due to generated content. Users in social networks establish *friendship* or *subscription* relationships with each other, and when users interact with or publish new content their *friends* are informed. We model these relationships with a *user graph* $\mathcal{G} = (U, E)$, where U is the set of users of the network and an edge $e = (u_i, u_j) \in E$ indicates that users u_i and u_j are friends (undirected case) or that user u_i follows user u_j (directed).

Moreover, a user may publish some new content item c_i, possibly *in response to* another content item c_j authored by another user, thus generating complex threads of discussion. Interactions within a single thread are modeled with a content *reply tree* $\mathcal{T} = (C, R)$, where C is the set of content items in the thread, and an arc $r = (c_i, c_j) \in R$ indicates that c_i is a reply to c_j. Note that $\mathcal{T}$ is indeed a tree as each content item, except the first one (the root), is a response to exactly one other item (its parent). Additionally, the nodes of $\mathcal{T}$ are enriched with information about publishing time and authoring user.

The tree $\mathcal{T}$ can be projected onto the users to model reply interactions among users. The resulting structure is a user *reply graph* $\mathcal{R} = (U, I)$,

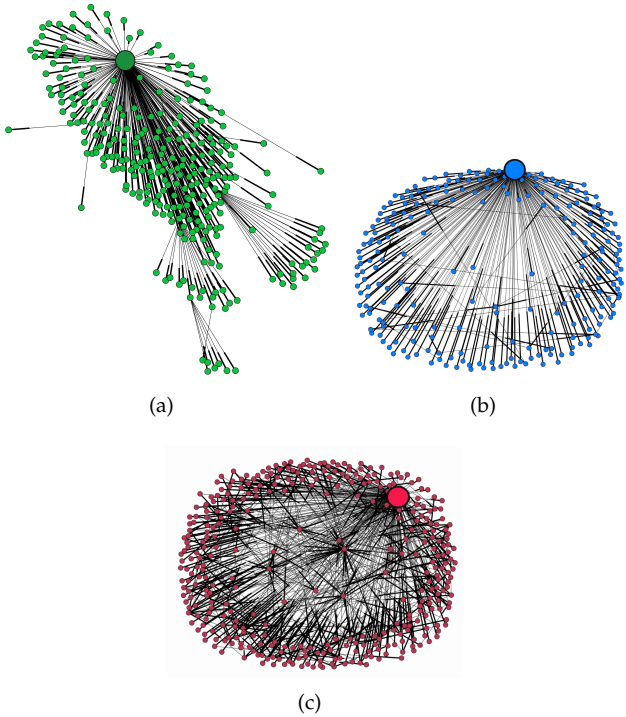


Figure 23: Examples of different user-interaction networks: (a) content reply tree; (b) user reply graph for a non-controversial topic; and (c) user reply graph for a controversial topic.

where an edge $e = (u_i, u_j) \in I$ indicates that the user u_i has replied to some content item posted by user u_j. We refer to the user who authored the first content item as *origin*.

Figure 6.23(a) shows a content *reply tree* (also referred to as just *reply tree*) present in our data, while Figure 6.23(b) and Figure 6.23(c) show the user *reply graph* (or just *reply graph*) of two other discussion threads. Note that a social network may have several disconnected reply trees and reply graphs.

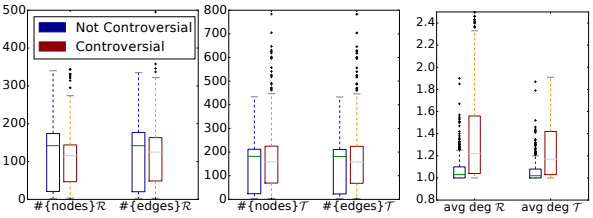
Our main hypothesis is that the structure of the user graph $\mathcal{G}$, the reply tree $\mathcal{T}$, and the reply graph $\mathcal{R}$ can be characterized by simple *motifs* of

local user interactions that can be effectively exploited to distinguish between *controversial* and *non-controversial* content.
In addition to local motifs, we also explore whether more traditional features (including network structure, content propagation, and temporal feautres) can be used to distinguish controversy. This standard graph-based analysis is discussed in Sec. 6.6.1, while the motif-based analysis is presented in Sec. 6.6.2.

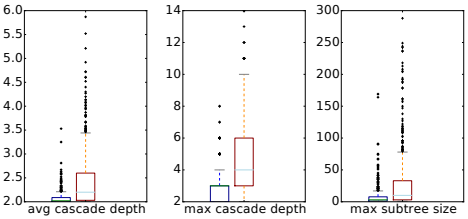
6.6.1 Standard graph-based analysis

Structural features. The simplest structural features to extract from the user-interaction networks are the *size* in terms of *number of nodes* and *number of edges*, and the *degree distribution*.
Figure 6.24(a) shows the distribution of the sizes of the reply tree $\mathcal{T}$ and the reply graph $\mathcal{R}$ in terms of number of nodes and number of edges for Dataset1 about Twitter pages with all the reply networks with at least 3 users involved in the conversation. To some extent, these measures are related to the popularity of the content taken into consideration. Note that in our data the sizes of $\mathcal{T}$ and $\mathcal{R}$ are very similar for both controversial and non-controversial content. This finding is somewhat surprising, as one would expect that controversial content generates larger threads of conversation. A plausible explanation is that our data-collection process favors popular topics (as such topics are more likely to be found in our crawl). Nevertheless, we can conclude that for distinguishing controversy *among popular topics*, just the graph sizes do not suffice.
Figure 6.24(b) reports the average degree for the reply tree $\mathcal{T}$ and the reply graph $\mathcal{R}$. In this case, the distributions are quite different for controversial and non-controversial content. A larger average degree is observed for controversial content, suggesting that such conversations generate more engagement among users.

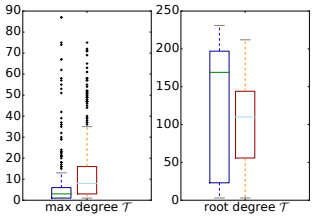
Propagation-based features. In order to understand how information propagates among controversial and non-controversial conversations, we investigate a number of different properties of the reply trees $\mathcal{T}$ related to



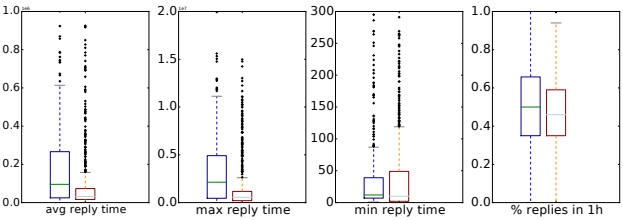
(a) Frequency (b) Structural



(c) Propagation I



(d) Propagation II



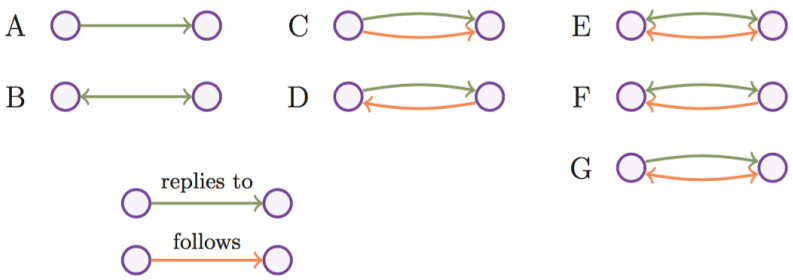
(e) Time

Figure 24: (a) Distribution of the number of nodes and edges in $\mathcal{T}$ and $\mathcal{R}$.
(b) Distribution of average node degree in $\mathcal{T}$ and $\mathcal{R}$.
(c) Distribution of avg./max. cascade depth and max. subtree size.
(d) Distribution of origin degree and max. degree in $\mathcal{T}$ and $\mathcal{R}$.
(e) Distribution of average, max., min. inter-reply time, and percentage of replies within one hour from the root. Non-controversial in blue (left side) vs. controversial in red (right side).

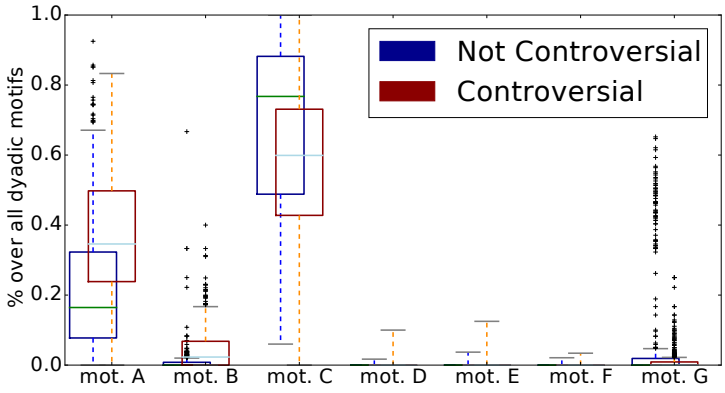
information propagation. Figure 24(c) shows the distribution of average and maximum cascade depths, where a cascade is defined as a path from the root to a leaf of a reply tree. The figure also shows the distribution of the maximum-size subtree among all subtrees rooted in a child of the root node. We observe that for controversial content the reply trees generally have larger depth.
Figure 24(d) reports the distribution of the degree for the root, as well as the node with the larger degree excluding the root in $\mathcal{T}$. We see that in this case the controversial and non-controversial discussions have similar distributions. Nevertheless, reply trees of controversial discussions have higher probability of having a smaller root degree than non-controversial, suggesting that controversial discussions go beyond the first level of interaction.
Given the above analysis, to summarize content propagation, we decide to use the two most significant features in the content reply trees. The other features, e.g., max cascade depth, are discarded because they are strongly related to popularity. In particular:

- *average cascade depth*: the average length of root-to-leaf paths;
- *maximum relative degree*: the largest node degree excluding the root node, divided by the degree of the root.

Temporal features. Considering the simple assumption that controversial topics may generate "dense" discussions in time, we analyze the time elapsed between a content item and its reply. Figure 24(e) shows the distributions of minimum, maximum and average inter-reply time. Additionally, we measure the ratio of nodes in a reply tree occurring within one hour from the root. For all the measures above, there is no significant difference between controversial and non-controversial reply trees. For prediction purposes, we choose to use as features only the average inter-reply time and the ratio of replies in the first hour. Maximum and minimum inter-reply time are influenced by a single reply and for this reason they were not considered further.



(a)



(b)

Figure 25: (a) Dyadic motifs and (b) their frequency distribution.

6.6.2 Motifs

Our main hypothesis is that *local patterns* of user interaction can be used to discriminate between controversial and non-controversial discussions. This hypothesis is consistent with previous studies, where it was shown that local patterns can be used to characterize different types of networks (BDGL08; MSOI+02). We consider local patterns to be 2- and

3-node connected subgraphs. We refer to such patterns as *motifs*.
We consider motifs in the user graph $\mathcal{G}$ and the reply graph $\mathcal{R}$. These two graphs encompass two different kinds of information. An edge in the user graph $\mathcal{G}$ indicates that a user is interested in the content produced by another user. These two users are likely to have similar interests and/or opinions. On the other hand, the reply graph $\mathcal{R}$ models the activity among users who may not know each other but are willing to discuss or comment on a specific topic. In this sense, the reply graph $\mathcal{R}$ is much more dynamic and content-dependent. Antagonism between users, which can not be captured by the user graph $\mathcal{G}$ can be captured by the reply graph $\mathcal{R}$. Our basic assumption is that a combined analysis of the two graphs, $\mathcal{G}$ and $\mathcal{R}$, can lead to an improved model for controversy detection.

Dyadic motifs. We consider all possible patterns between two users in graphs $\mathcal{G}$ and $\mathcal{R}$, such that that there is at least one reply (i.e., one edge in graph $\mathcal{R}$) — otherwise the two users do not interact with each other in the discussion thread. There are seven possible configurations, which are shown in Figure 25(a). Figure 25 shows the frequency distribution of dyadic motifs in our data. Note that patterns are mutually exclusive, therefore, pattern A where u_i replies to u_j also implies than u_j does not reply u_i and that the two users do not follow each other.
The most frequent dyadic motifs are A and C. According to Figure 25, it is more likely to observe a reply to a followed user in non-controversial cases. Conversely, in controversial cases it is likely to reply to a user not being followed. This confirms the intuition that controversial discussions thread interactions also among users not directly connected in the user graph $\mathcal{G}$.
The features used for detecting controversial content are the frequencies of all dyadic motifs.

Triadic motifs. We also consider 3-node motifs, in particular closed triangles. As in the case of dyadic motifs, we combine structural information from the user graph $\mathcal{R}$ and the reply graph $\mathcal{G}$. Figure 26(a) shows some master motifs we considered. We again consider motifs only

if there is a reply interaction among the three users. Due to the high number of possible motifs and since most motifs are relatively rare in the data, we coalesce motifs in groups. Overall, we form our set of triadic motifs by considering (i) the number of follow edges among the three users (Figure 26(a)), (ii) the number of reciprocal follow edges, and (iii) the number of non reciprocal follow edges with opposite direction with respect to the reply edge. In total we have 20 different triadic motifs. The frequency of each motif is considered as a feature for predicting controversy.

We do not report the distribution for all the motifs, but generally most of the patterns we considered for closed triangles were quite rare in the dataset. Only a few of them are frequent and mostly in controversial threads, confirming the intuition that controversial discussions exhibit a more complex structure.

To provide additional insights on user interactions, we consider as additional feature the ratio of triangles in the reply graph $\mathcal{R}$ over the number of all possible triangles $\binom{|U|}{3}$. Again, a larger triangle ratio indicates that controversial content generates more complex discussion threads with more interactions among users and not only dyadic relations between the author of the post and the replying user, as it in the case of non-controversial situations.

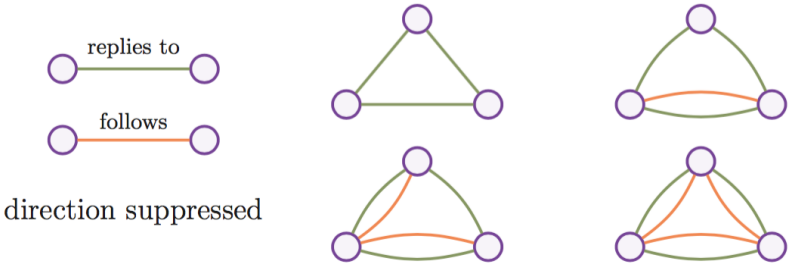
We also considered "open" triadic motifs, i.e., 3-user subgraphs connected by only two replies. Such patterns did not seem to help much in predicting controversial discussions and therefore they are not considered further.

In summary, all features used for the task of predicting controversial content are shown in Table 20.

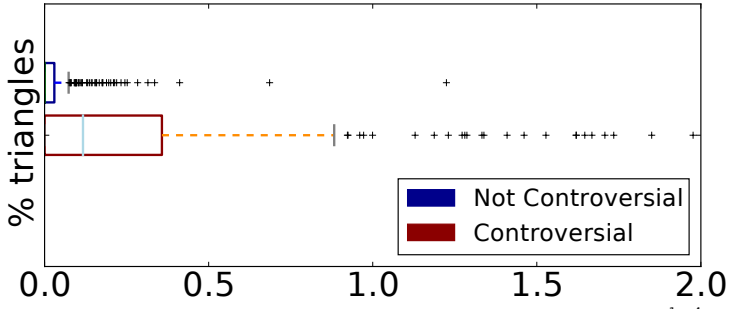
6.7 Evaluation

6.7.1 Detection of controversy in Twitter pages

We used the Twitter datasets presented in Section 6.5. As already discussed, the Twitter pages of Dataset1 can be considered entirely controversial or non-controversial, therefore we labeled each tweet according to the



(a)



(b)

Figure 26: (a) Triadic motifs and (b) distribution of undirected reply triangles ratio.

page it belongs. The dataset is quite balanced, with about 60% instances belonging to the controversial class and 40% to the non-controversial. Reported experiments are performed using 5-fold cross-validation and averaged over 100 trials.

We evaluated different classifiers, including AdaBoost, Logistic Regression, SVM, Random Forest, etc., and chose AdaBoost as it resulted in the best performance. We analyzed the performance by the standard graph-based features and by additionally using motif-based features. We report the accuracy of the classifier on both controversial and non-

Table 20: Summary of all features

structural	avg degree in $\mathcal{T}$ avg degree in $\mathcal{R}$
propagation	avg cascade depth in $\mathcal{T}$ max degree in $\mathcal{T}$ / root degree in $\mathcal{T}$
temporal	avg inter-reply time % replies in 1h
dyadic motifs	7 2-node motifs (shown in Figure 25(a))
triadic motifs	20 3-node motifs triangles ratio

Table 21: Performance of motif-based classifier

Filtering	Accuracy	Precision	Recall	F-measure
structural, propagation-based and temporal features only				
$>$ 2 users	0.76	0.79	0.81	0.80
$>$ 3 users	0.77	0.80	0.82	0.81
$>$ 10 users	0.78	0.81	0.83	0.82
with addition of dyadic motifs				
$>$ 2 users	0.82	0.84	0.86	0.85
$>$ 3 users	0.83	0.85	0.86	0.85
$>$ 10 users	0.84	0.86	**0.88**	**0.87**
with addition of triadic motifs				
$>$ 2 users	0.83	0.85	0.86	0.85
$>$ 3 users	0.84	0.86	0.85	0.86
$>$ 10 users	**0.85**	**0.87**	**0.88**	**0.87**

controversial classes, and the precision, recall and F-measure with respect to the controversial class.

As shown in Table 21, when using structural, propagation-based and temporal features only, accuracy is above 75% and increases only slightly when restricting to reply trees with more than 10 users. With the addition of dyadic motifs, all the performance figures are significantly improved. Note that the precision of the algorithm improves in both controversial and non-controversial classes.

The addition of triadic motifs leads to the best results, but the improvement is only marginal. This is because, as discussed in Section 6.6, triads

are infrequent: even if conveying relevant information, they may help in improving the classification of a limited number of instances.
In Table 22 we reported the 8 most relevant features exploited by the AdaBoost model according to the error reduction. Temporal features are important to detect controversy. The first feature is the average inter-reply time, and the fourth is the ratio of replies posted within one hour of the original tweet: when the discussion is polarized people tend to reply in a shorter time. This result is in line with other contexts. For example, it is known that temporal features play the main role to predict popularity (SSC16). The second most important feature is the maximum relative degree, i.e., the maximum degree normalized by the root node degree. In non-controversial reply trees, the root is the only node with a large degree, i.e., the node attracting most of the reply activity.
The other features among the top-6 are dyadic motifs. The most relevant being motif A which corresponds to a user u_i replying to u_j without any following relationship among the two. We deduce that controversial threads create engagement among users not being directly connected in the social network. On the other hand, the non-relevancy of motif C (where a user replies to a follower), suggests that it is less likely to have controversial discussions among friends. Interestingly, dyadic patterns resulted being more relevant than propagation-based features. For instance, the depth of the cascades, which was expected to model the complexity of the interactions, is not among the top-8 features. Presumably, complex propagation features are superseded by the simple motif patterns.
Finally, the last two important features are based on triangles. In particular the relevance of the triangle-ratio feature suggests that triadic patterns are able to grasp interactions occurring in controversial discussions. However it is harder to draw any conclusion on the role of specific triads patterns, due to their low frequency. The most significant specific triadic pattern included in the list in Table 22 is a close reply triangle with two follow edges: one reciprocal and one not reciprocal with the same direction of the underlying reply edge. Since triadic patterns provide a limited contribution to the classifier, we conclude that dyadic motifs are already effective, and there is not much information that can be extracted based

Table 22: Feature importance (filtering > 10 users)

Feature	Error Reduction
Avg. inter-reply time	0.18
Max. relative degree	0.16
Motif A	0.14
% Replies within 1h	0.08
Motif B	0.08
Motif G	0.06
Triangles ratio	0.04
Triadic motif	0.04

on specific triadic motifs.

6.7.2 Dynamic tracking of controversy

We found it is not always appropriate to classify a reply tree as controversial or not. This is because each reply may generate unexpected reaction. For instance, there may be sub-threads of controversy, within a non-controversial discussion. To test this intuition, we analyzed the direct replies of the *origin* tweets that were classified as non-controversial. This can be achieved easily as the proposed approach can be applied to any tweet given its reply tree, or in this case, its reply sub-tree. By applying the model discussed in the previous section, we found that about 7% of the direct-reply sub-trees of a non-controversial tweet are controversial. One example is shown in Figure 27, illustrating the reply tree of a post by Justin Bieber. The majority of the replies are not controversial and are written by his fans with compliments and expressions of affection and love. However, the proposed algorithm detected as controversial one sub-tree (highlighted in red) generated by a reply in support to another singer: "Zayn is better". This post generated a subtree with animated discussion among fans. A similar case was found for Cristiano Ronaldo's profile, where a number of users started discussion about his rivalry with Messi.

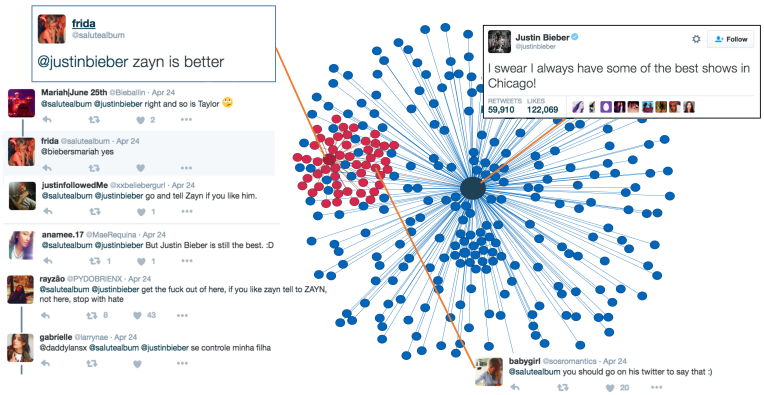


Figure 27: A controversial reply sub-tree (red) originated by a non-controversial post (blue) by Justin Bieber

Both of the previous examples are typical cases in which the controversial portion of the discussion is limited to a few branches, and its detection might be challenging. We claim that the proposed approach, based on local motifs can successfully detect small controversial sub-threads.

6.7.3 Hashtags evaluation

Since on Twitter, topics are often identified through hashtags, we tested the proposed method on tweets mentioning a given hashtag (Dataset2), obtained from the previous work (GDFMGM16). Table 23 shows the fraction of controversial posts per hashtag, as detected by our model. The smallest fraction of controversial discussions is found with #sxsw and #ultralive hashtags (related to music events), where most conversations are expected to happen among supporters of the same music band. The most controversial discussion are found with the #beefban, #onedirection, #netanyahu, #baltimore hashtags. The classification of the these hashtags as controversial is in line with the previous results (GDFMGM16), with the exception of #onedirection for which we detected antagonist replies, upon manual inspection. Most of the hashtags exhibit a mixed behavior

Table 23: Hashtag controversy classification

Hashtag	Ratio of controversial posts
sxsw	0.32
germanwings	0.49
beefban	0.70
netanyahu	0.55
ultralive	0.29
onedirection	0.61
baltimore	0.58
russia-march	0.46

as far as controversy is concerned.[1]

Indeed, simply counting the number of tweets classified as controversial is a quite naïve approach, strongly dependent on different factors, such as the daily volume of tweets, on external events, and many others. For these reasons, we believe that it is more interesting to study how the controversy related to a given hashtag evolves over time.

Figure 28 shows the evolution of the controversy for the #germanwings hashtag. Note that some hours after the accident happened on March 24 the majority of threads is controversial. In the evening the discussions become less controversial and mainly about sorrow and condolences. An interesting increase of the controversy level is registered the next day, until details about the accident were released. Then the discussion becomes predominately non-controversial showing that the news has been digested by the audience. We highlight that the level of controversy is anti-correlated to the frequency for motif A, thus confirming the prediction power of the proposed motifs.

[1]For example: A controversial tweet from #germanwings, `https://twitter.com/stephaneguillon/status/580330769912061953` and a non-controversial tweet about #baltimoreriots `https://twitter.com/jelani9/status/592102034935013376`

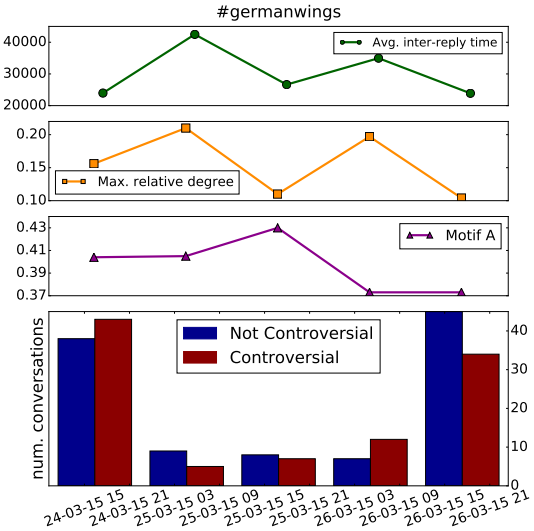


Figure 28: Distribution of controversial (red) vs. non-controversial (blue) posts and top-3 features values over time for the #germanwings hashtag

6.8 Conclusion

We proposed a novel approach based on local graph motifs for controversy analysis in OSNs. The proposed method is language independent and exploits local patterns of user interactions to detect controversial threads of discussion.

Given a content item, users reply to each other generating different configurations of the reply graph. We investigated local motifs extracted from this graph and from the user friendship graph. Such motifs correspond to different interaction patterns among two users, which may be linked by a possibly reciprocal reply action and by a possibly reciprocal friendship relationship. Similar motifs regarding the interaction of three users were considered.

We proved on a benchmark Twitter dataset that such motifs are more

powerful in predicting controversy than other frequently-used graph properties such as cascade depth. We observed that in most cases controversy arise when users participate to discussions beyond their polarized communities.

Finally, as the proposed motifs can be easily extracted from any reply tree or sub-tree, we experimented with the use of such patterns in monitoring the evolution of discussions and sub-discussions over time. Indeed, we found that a topic of discussion develops over time changing its level of controversy depending on different sub-topics or on external events (e.g., news). Therefore, a fine-grained analysis, as provided by the proposed local motifs, is necessary for a better understanding of controversy in online social networks.

Chapter 7

Content diffusion: deviant communities behavior

The results discussed in this chapter were published in (CALS16a) and submitted to (CALS16b).

- Coletto, M., Aiello, L. M., Lucchese, C., Silvestri, F. (2016). On the Behaviour of Deviant Communities in Online Social Networks. In 10th International AAAI Conference on Web and Social Media. ICWSM 2016, May 17-20, Cologne, Germany.
- Coletto, M., Aiello, L. M., Lucchese, C., Silvestri, F. (2017). Adult Content Consumption in Online Social Networks. Submitted to Social Network Analysis and Mining, edited by Reda Alhajj (Springer).

7.1 Introduction

In the previous chapters we described Online Social Networks (OSNs) as complex ensembles of inter-linked *polarized communities* that interact on different topics. Some communities are well connected among each other through controversial discussions (e.g., political party supporters in Twitter), while in other cases they are segregated, with most of the interactions among members rather then with external communities and

the rest of the social network. In other cases the segregation of the community might be related to the topic of interaction that can commonly be considered inappropriate with respect to the society's norms or moral standards.

Those are communities that depict or discuss what are usually referred to as *deviant behaviors* (CM15), conducts that are commonly considered inappropriate because they are somehow violative of society's norms or moral standards. Pornography consumption, drug use, excessive drinking, eating disorders, or any self-harming or addictive practice are all examples of deviant behaviors. Many of them are represented, to different extents, on social media (HIJW10; MSEB10; DC15). However, since all these topics touch upon different societal taboos, the common-sense assumption is that they are embodied either in niche, isolated social groups or in communities that might be quite numerous but whose activity runs separately from the mainstream social media life. In line with this belief, research has mostly considered those groups in isolation, focusing predominantly on the patterns of communications among community members (TESU15) or, from a sociological perspective, on the motivations to that make people join such groups (Att05).

In reality, people who are involved in deviant practices are not segregated outcasts, but are part of the fabric of the global society. As such, they can be members of multiple communities and interact with very diverse sets of people, possibly exposing their deviant behavior to the public.

In this chapter we focus on adult content consumption networks, which are present in many on-line social media and in the Web in general. We found that a few small and densely connected communities are responsible for most of the content production. Differently from previous work, we study how such communities interact with the whole social network. We found that the produced content flows to the rest of the network mostly directly or through bridge-communities, reaching at least 450 times more users. We also show that a large fraction of the users can be inadvertently exposed to such content through indirect content resharing.

7.2 Problem formulation

In this chapter we aim at answering the following research question:
Q6: *How does content spread beyond niche or segregated communities?*
Even though in many contexts users are segregated in their *echo chambers* it might be that the produced content spreads through the weak ties or through the controversial interactions and goes beyond the producer communities.
The current chapter is dedicated to the study of deviant communities, formation of topical communities centered on matters that are not commonly taken up by the general public because of the embarrassment, discomfort, or shock they may cause. These are polarized communities or at least topical communities usually considered very isolated from the rest of the social network. Since all these topics touch upon different societal taboos, the common-sense assumption is that they are embodied either in niches or in communities that might be quite numerous but whose activity runs separately from the mainstream social media life.
We show that for specific deviant communities even though the producers are a small group, the content spreads far from the members who created it. Our analyses have been performed on Tumblr and on Flickr.

7.3 Contribution

In the work presented in this chapter we aim to go beyond previous studies that looked at deviant groups in isolation by observing them *in context*. In particular, we want to shed light on three matters that are relevant to both network science and social sciences: *i)* how much deviant groups are structurally secluded from the rest of the social network, and what are the characteristics of their sub-groups who build ties with the external world; *ii)* the extent to which content produced by a deviant community spreads and is accessed (voluntarily or inadvertently) by people outside its boundaries; and *iii)* what is the demographic composition of producers and consumers of deviant content and what is the potential risk that young boys and girls are exposed to it.

In this initial study we undertake to answer those questions focusing on the behavior of *adult content* consumption. Public depiction of pornographic material is considered inappropriate in most cultures, yet the number of consumers is strikingly high (SWF08). Despite that, we are not aware of any study about the interface between adult content communities and the rest of the social network. We study this phenomenon on a large dataset from Tumblr, considering big samples of the follow and reblog networks for a total of more than 130 million nodes and almost 7 billion directed dyadic interactions. To spot the community that generated adult content, we also recur to a large sample of 146 million queries from a 7-month query log from a very popular search engine (Section 7.5), out of which we build an extensive dictionary of terms related to adult content that we make publicly available.

Results show that:

- The deviant network is a tightly connected community structured in subgroups, but it is linked with the rest of the network with a very high number of ties (Section 7.6.1).

- The vastest amount of information originating in the deviant network is produced from a very small core of nodes but spreads widely across the whole social graph, potentially reaching a large audience of people who might see that type of content unwillingly. Although the consumption of deviant content remains a minority behavior, the average local perception of users is that neighboring nodes reblog more deviant content than they do (Section 7.6.2).

- There are clear differences in the age and gender distributions between producers and consumers of adult content. The differences we found are compatible with previous literature on adult material consumption: producers are older and more predominantly male and age greatly affects the consumption habit, strengthening it in males and weakening it in females (Section 7.6.3).

7.4 Related work

Groups in online social media. Computer science research has dealt extensively with the problem of classification of groups along structural, temporal, behavioral, and topical dimensions (NGP08; GAEJ13; Aie15). The relationship between group connectivity and shape of information cascades has also been explored, revealing an intertwinement between community boundaries and cascade reach that is particularly tight in communities built upon a common theme shared by all of their members (EK10; RTU13; BBM13; MBAG$^+$14b). The degree of inter-community interaction has been analyzed mostly in the context of heavily polarized networks, the most classical example being online discussions between two opposing political views (AG05b; CRF$^+$11a; FKSW11). These studies explored methods to quantify segregation (GMJCK13), but mainly focus on networks formed by two main divergent clusters.

Deviant communities. Deviant networks have been analyzed mostly in isolation. Studies about the depiction of drug and alcohol use in social media adopted mainly the content perspective. Researchers aimed at identifying the elements that boost content popularity, investigated the effect of gender on engagement, and studied the perceptions that deviant content arises in the young public (MSEB10). Research has been conducted around anorexia-centered online communities (GRP08; RPNB11; BP12), also on Tumblr (DC15), investigating a wide range of aspects including the construction and management of member identities, the processes of social recognition, the emergence of group norms, and the use of linguistic style markers. Similar studies have been published over the years on communities of self-injurers and negative-enabling support groups, in which members encourage negative or harmful behaviors (HIJW10). Fewer studies touch upon network-related aspects. One notable example is the work by Tyson et al. (TESU15) that provides an overview of behavioral aspects of users in the PornHub social network, with particular focus on the role of sexuality and gender. More loosely related are studies on the so-called *dark networks*, mostly motivated by the need of finding

effective methods to disrupt criminal or terroristic organizations (XC08). The study by Christakis et al. (CF08) about the communication network between smokers and non-smokers is one of the few quantitative studies that addresses the interaction between the social network and one of its sub-groups, but it strongly focuses on the phenomenon of contagion.

Adult content consumption. In the context of internet pornography consumption, computer science literature studied the categorization of content and frequency of use (SZV13; TESU13; HŠ15). A wider corpus of research has been produced by social and behavioral scientists by means of surveys administered to relatively small groups. Special attention has been given to the relationship between age or gender and the exposure (voluntary or unwanted) to internet porn (SWF08; YM05; Buz05; MFW03; CLCY13), with particular interest to the age band of young teens (MFW03; CLCY13; WMF07). Numbers vary substantially between studies, but clearly men are more exposed than women (approximately 75%-95% vs. 30%-60%), with men exposed more frequently (Hal06) and women more often involuntarily. It is estimated that young teens that are often exposed accidentally (roughly 25% to 66% of the times) and are also exposed to violent or degrading pornography (20% among female, 60% among male) (RB15). Researchers have also pointed out the potential harm that adult material consumption through internet can cause, including addiction (KG14) and increased chance of adopting aggressive behavior (ADB95). Exposition also correlates with drug use (YM05) and with lack of egalitarian attitude towards the other sex (HML13). Although delving into the potential harm of pornography is far beyond the scope of our work, this inherent risks provide an additional motivation to focus on this particular type of deviant community.

7.5 Data

This study uses data collected from Tumblr, a popular micro-blogging platform and social networking website. The dynamics of the Tumblr community are based mostly on three possible actions. Users can *post* new

entries on their blogs usually containing multimedia content, *repost* on their blogs any post previously published by others (similarly to Twitter retweets), and *follow* other users to receive updates from their blogs in a stream-like fashion. Users might own multiple blogs, but for the purpose of this study we consider blogs as users, and we will use the two terms interchangeably.

We consider as *deviant nodes* those users who post content about a given *deviant topic*. To identify deviant nodes we resort to data from search logs. As shown in other studies (LC11), if a deviant query *hits* (i.e., leads to the click of) a Tumblr blog URL, then the blog is a candidate *deviant node*.

In our analysis we use a seven-month long query log (from Jan. to Jul. 2015) of a major search engine, from which we collected a random sample of 146M query log entries whose clicked URL belongs to the `tumblr.com` domain. We limit our study to queries that were submitted from the United States. After a simple query normalization process involving lowercasing and the removal of numbers, additional spaces, and of the word "tumblr" with its most common misspellings (as observed from the term distribution) we obtained about 26M unique queries that hit a total of 2.7M unique Tumblr blogs. As expected, the distribution of number of queries hitting a blog is very skewed, with most popular blogs being reached by hundreds of thousands of clicks originating from search queries (Figure 29). In the remainder of the chapter, we focus on adult content, this being a very common *deviant* topic on the Web. The same kind of analysis could be conducted on any other *deviant* topic.

To maximize the accuracy and coverage of the set of discovered deviant nodes, we devise an iterative semi-supervised *Deviant Graph Extraction* procedure. Given a query log $\mathcal{Q}$, and a set $\mathcal{K}_i$ of deviant keywords (possibly multi-grams), we define as $\mathcal{Q}(\mathcal{K}_i)$ the set of queries in $\mathcal{Q}$ that exactly match any of the keywords in $\mathcal{K}_i$. Based on the query log information, the set $\mathcal{Q}(\mathcal{K}_i)$ yields a collection of clicked URLs from which we selected those corresponding to blogs in the Tumblr domain. We denote such set of blogs as $\mathcal{B}(\mathcal{K}_i)$. To reduce data sparsity, we filter out the blogs in $\mathcal{B}(\mathcal{K}_i)$ with less than two unique incoming queries in $\mathcal{Q}(\mathcal{K}_i)$ or less than 3 clicks originated by them.

The set of queries hitting $\mathcal{B}(\mathcal{K}_i)$ is used to create a new set of keywords $\mathcal{K}_{i+1}$ and to re-iterate the procedure. Given the current set of deviant nodes $\mathcal{B}(\mathcal{K}_i)$ we identify the 10% of blogs with highest proportion of query hits that match words in $\mathcal{K}_i$; those are the blogs that are hit mostly by deviant queries compared to other query types. We select all the unique queries that hit those blogs and merge them with $\mathcal{K}_i$, thus obtaining a new set of keywords $\mathcal{K}_{i+1}$, which is used to feed the next iteration of the algorithm. The procedure is repeated until the sizes of both $\mathcal{K}_i$ and $\mathcal{B}(\mathcal{K}_i)$ converge.

The initial set $\mathcal{K}_0$ is obtained as follows. We first create a keyword set as the union of the search keywords from professional adult websites along with the list of adult performers published by movie production companies. To extend the coverage also to blogs that are reached predominantly by Spanish queries (the second most used language in US), we also translated to Spanish the initial set of keywords. From this initial set we manually extracted two dictionaries of respectively $5{,}152$ and $5{,}283$ search keywords (mono-grams, bi-grams, multi-grams), which were used to filter queries in the query log following two strategies: 1) *exact match*, selecting those queries in the query log which match exactly one search keywords in the first dictionary, 2) *containment*, selecting those queries subsuming any search keywords term in the second dictionary. For instance, the word *porn* is not included in the containment dictionary because queries like *food porn* should not to be detected as adult. The union of the queries detected by the two strategies hits a set of blogs, whose most frequent incoming queries were manually inspected to detect further 351 search keywords. The union of these terms with the *exact match* dictionary leads to a set of $5{,}503$ *deviant queries* ($5{,}152$ + 351) which is used as the seed set $\mathcal{K}_0$ to bootstrap the *deviant graph* extraction.

The above algorithm is biased towards the query log data, and on the popularity of blogs measured through the volume of search queries. On the other hand, this method allows to identify very quickly nodes that are likely to be relevant in the network as they produce the most interesting content to Web users. Also, as the procedure is network-oblivious (the graph structure is not exploited), no bias is introduced in our analysis of

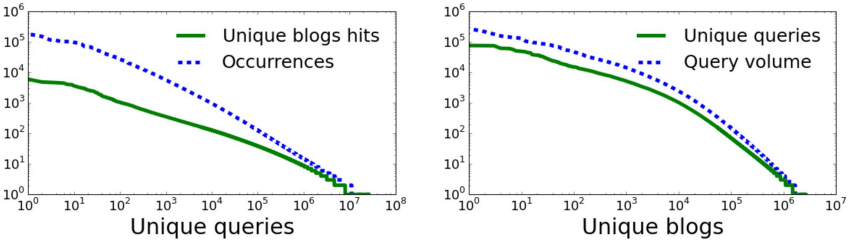


Figure 29: Distributions of: (left) number of blogs hit by a query and number of occurrences of a query; (right) volume of (unique) queries hitting a blog.

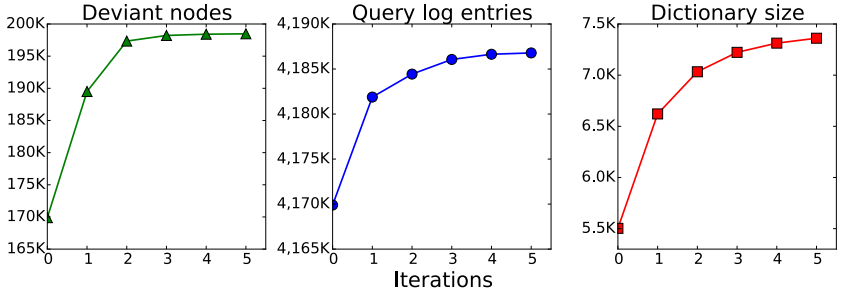


Figure 30: Convergence of the three quantities used in the Deviant Graph Extraction procedure

the network.

Figure 30 shows that the *Deviant Graph Extraction* procedure converges quickly. We stop after 6 steps with $198K$ nodes hit by $4.2M$ unique queries. The final vocabulary containing 7,361 words is made publicly available to the research community[1]. In Figure 31 we report the distribution of the *deviant query volume* ratio for the deviant nodes detected. The distribution is skewed, showing that about 30% of the nodes are hit by a majority of deviant queries.

To study the interaction of deviant nodes with the rest of the social network, we extracted a subset of the Tumblr follower and reblog networks

[1] `https://github.com/hpclab/DevCommunities/`

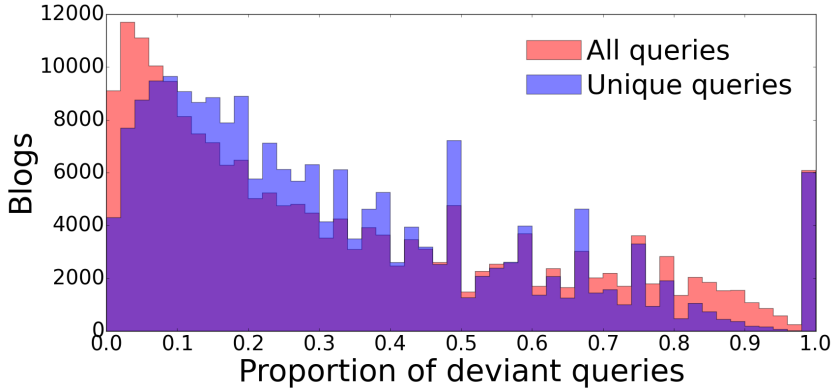


Figure 31: Distribution of deviant query volume ratio reaching deviant nodes

with a snowball expansion starting from the $198K$ identified deviant nodes up to 3-hops away. The follower is a snapshot of the graph done in December 2015; the reblog network was built from the reblog activity happened in the same month. Statistics about the resulting networks are reported in Table 24.

We also obtained information about self-declared age and gender for about $1.7M$ Tumblr users and, in particular, for about 10% of the detected *deviant nodes*. The datasets include exclusively interactions between users who voluntarily opted-in for such studies. All the analysis we report next has been performed in aggregate and on anonymized data.

7.6 Analysis

The availability of data about the interaction between deviant nodes and the social network that surrounds them provides the unique opportunity to study the structure and dynamics of a deviant network within its context. We first analyze the shape of the deviant network and measure its connectivity with the rest of the social graph (Section 7.6.1). We then look into how the information originating from deviant networks spreads across the boundaries of the deviant group (Section 7.6.2). Last, we study

Table 24: Tumblr - Network statistics for the reblog (R) and follow (F) networks of the full graph sample (*All*), the deviant graph (*Deviant*), and the four communities that compose it (*Producers*$_{1,2}$ and *Bridge*$_{1,2}$). All the statistics are about the giant weakly connected components and count only links whose both endpoints are in the considered node subset. $\langle k \rangle$=average degree, D=density, ρ=reciprocity, C=clustering, $\overline{spl}$=average shortest path length, d=diameter.

	$\|N\|$	$\|E\|$	$\langle k \rangle$	D	ρ	C	$\overline{spl}$	d
All R	14M	472M	33	$2{\cdot}10^{-6}$	0.06	-	-	-
All F	130M	6,892M	53	$4{\cdot}10^{-7}$	0.10	-	-	-
Deviant R	105K	1.4M	13	$1{\cdot}10^{-4}$	0.04	0.10	3.73	11
Deviant F	135K	24.6M	182	$1{\cdot}10^{-3}$	0.07	0.13	2.80	8
Prod$_1$ R	48K	914K	19	$4{\cdot}10^{-4}$	0.04	0.09	3.44	9
Prod$_2$ R	16K	305K	19	$1{\cdot}10^{-3}$	0.05	0.13	3.19	8
Bridge$_1$ R	9K	36K	4	$5{\cdot}10^{-4}$	0.04	0.08	4.18	13
Bridge$_2$ R	3K	32K	11	$4{\cdot}10^{-4}$	0.06	0.21	3.32	10

some demographic properties that characterize producers and consumers (Section 7.6.3).

7.6.1 Deviant network connectivity

The deviant network is a tiny portion of the whole graph, representing about 0.7% of all the nodes in the reblog graph and 0.1% of those in the follow network. So few nodes could be scattered along the social network or clustered together. So we ask:

AQ1: *Are deviant nodes organized in a community?*

We consider the deviant networks as the subgraphs of the follow and reblog Tumblr networks induced by the *deviant nodes*. A directional link in the follow (reblog) network from node i to node j exists if i follows (or reblogs the posts of) j, meaning that the information flows from j to i. Basic network statistics on such subgraphs reveal that the deviant networks are quite dense, yet they have a high diameter (Table 24). Similar statistics have been observed before in other social networks (ABC$^+$10) and might be an indication of the presence of strong sub-groups patterns, as well as a signal of the absence of a community structure. To better

determine the reason for such elongated shape, we run the Louvain community detection algorithm (BGLL08b) on the deviant network[2]. Four clusters emerge, whose network statistics are summarized in the bottom lines of Table 24. To determine their nature, we manually inspected the content of 250 blogs in each of them.

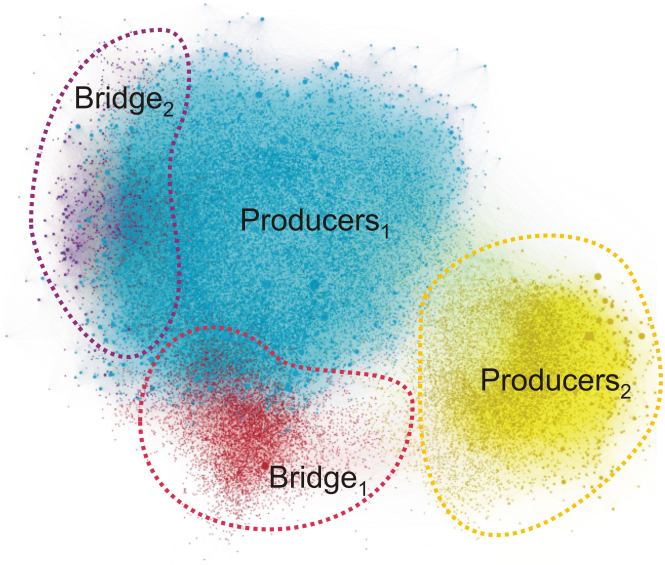


Figure 32: Bird-eye view of the deviant network in Tumblr, with colors denoting algorithmically-extracted communities

More than 90% of all the blogs in the two largest clusters contain blogs that *exclusively* produce explicit adult content, aimed at an heterosexual public (*Producers*$_1$) or at a male homosexual public (*Producers*$_2$). The blogs in the two remaining communities post less explicit adult content and

[2]Louvain is a modularity-based graph clustering algorithm that shows very good performance across several benchmarks (For10) and that is fast to compute even on large networks.

more sporadically, often by means of reblogging. They either focus on celebrities ($Bridge_1$), or function as aggregator blogs with high content variety, including depiction of nudity ($Bridge_2$).
From a bidimensional visualization of the network layout (Figure 32) it becomes apparent that the two bigger clusters are two well-separated cores that give a characteristic hourglass shape to the network, reason for the high diameter observed. The remaining communities are peripheral and arranged in a crown-like fashion (which explains their high diameter) around the largest sub-cluster $Producers_1$. We name the two smaller groups *bridge communities* as their main focus is not on deviant content but they are an entry point for deviant query traffic and, as we shall see next, act also as bridges towards the rest of the graph.
In short, we find that deviant nodes are not scattered in the social network but are tightly organized in a structure of distinct communities. To find out about the nature of their interaction with the rest of the social ecosystem, we proceed to answer the next question.

AQ2: *To what extent is the deviant graph connected to the rest of the social network?*

There are several ways to estimate the connectivity between two sets of nodes in a graph. We use different metrics to measure it between the four communities of the deviant network and the rest of Tumblr, as summarized by the matrices in Table 25; rows represent the group of nodes from which the social tie originates, columns those on which it lands.
The average volume of connections (Table 25, left) provides a first indication about the difference in connectivity across different groups. The diagonal has the highest values because of the community structure of the deviant network and of its sub-communities: members of a group have many more ties towards other group members rather than to the outside. This is true in particular for the two *Producer* clusters. The volume of links incoming to the largest producer cluster is particularly high from the smallest bridge community ($Bridge_2$), which surrounds it. The average Tumblr user in our sample follows around 51 users, between 2 or 3 of

which are in the core of the deviant network and around 2 of them are in bridge communities; similarly, among the 33 users reblogged in one month by the average user, one is from a *Producer* cluster and one from a *Bridge* group.
When looking at raw volumes, the amount of links from the deviant network to the rest of the graph is very high, mainly due to the high dimensionality of the set of nodes that are not deviant. To partially account for dimensionality of the groups, we measure the connectivity with density computed as the ratio of edges between the two groups over the total number of possible edges between them (Table 25, center). Also in this case the overall patterns hold, but the connectivity towards the external graph drops significantly.
Values of density are still affected by size, though. It is known that in real networks there is a strong correlation between density and number of nodes (LKF05). To fix that, in the spirit of established work in complex systems (SBC$^+$10) we resort to a comparison of the real network connectivity with a *null model* that randomly rewires the links while keeping the degree of each node unchanged. The values we report in Table 25 (right) indicate how many times the number of connections observed deviate from the null model. Also in this case, values on the diagonal are very high (except for the outer network, which has a value close to 1, as expected). Also, this computation highlights that ordinary users have a tendency to reblog content from the core of the deviant network almost 7 times more than random and between 16 and 53 times more than random from the bridge community members.
In summary, the core of the deviant community is dense but it is far from being separated from the rest of the graph, which is connected to it both directly and even more tightly through bridge groups.

7.6.2 Deviant content reach

We found that, although the deviant network forms a tightly connected community, it is not isolated from the rest of the social graph. This calls for an investigation about the visibility that the deviant content has in the

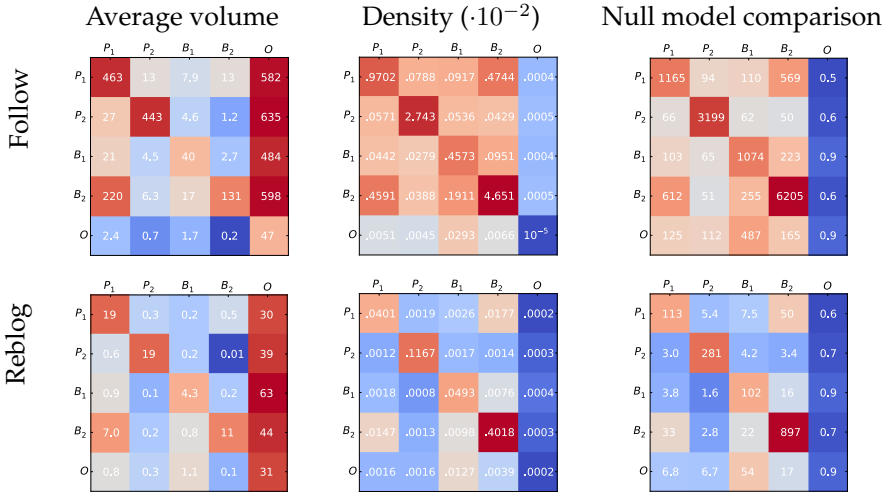


Table 25: Measures of connectivity between the communities in the deviant network (*Producers* P_1, P_2 and *Bridges* B_1, B_2) and the rest of the social network O, for both the follow (top) and reblog (bottom) relations. Link directionality is considered: ties originate from groups listed on the rows and land on groups listed on the columns.

outer network and what are the main factors that determine its exposure. We do so by answering the three research questions below.

AQ3: *How much deviant content spreads in the social graph and who are the main agents of diffusion?*

The exposure to deviant content goes beyond the members of the deviant network who are the *producers* of original adult material. Specifically, the *consumers* of deviant content can be categorized in three classes. The first is the class of *active consumers*: nodes who reblog (but not necessarily follow) adult posts, thus contributing to its spreading along social ties. Posts can be re-blogged in chains and create diffusion trees that potentially spread many hops away from the original content producer, therefore active consumers could further be partitioned in those who spread the content *directly* from the producers and those who do it with *indirect* reposts. The

second is the class of *passive consumers*: nodes who do not contribute to the information diffusion process but are explicitly interested in adult content because they directly follow the producer nodes. The last class is the one of *involuntary consumers* (or *unintentionally exposed* users): users who do not follow any producer node and do not reblog their content, but happen to follow at least one active consumer who pushes adult content in their feed through reblogging.

By drawing a quantitative description of the volume of deviant content reaching these three classes we can estimate how much the adult community is visible in the network at large. We adopt a conservative approach in which we consider the two *Producers* communities as the only ones generating original explicit (homosexual and heterosexual) content. Given the results of the aforementioned manual inspection, we are very confident that their activity is completely focused on the production of adult material.

We measure the size of the different consumer classes and the amount of content that flows through or to them by means of reblogging. The results are summarized by the schema in Figure 33. The network of deviant content producers is very small but receives a considerable amount of attention from direct observers. The audience of passive consumers counts almost 24M people. Around 2M users reblog directly from the deviant network, for a total of around 28M reblog actions in one month. A consistent part of the two *Bridge* communities within the deviant graph (a total of 3K users) are also direct consumers, and they reblog *Producers* 56K times per month. When looking at the set of 2.4M users who indirectly reblog deviant content, we see that only a small fraction of their monthly reblogs (less than 7%) is performed through bridge communities. However, in relative terms, bridge communities are considerably more efficient in spreading information than the average active consumer. If we consider efficiency η of a user set U as the ratio between reblogs done r_d and reblogs received r_r, weighted by the cardinality of the set $\left(\eta = \frac{r_r}{r_d \cdot |U|}\right)$, we discover that the bridge communities ($\eta = 1.5 \cdot 10^{-3}$) are several orders of magnitude more effective in spreading the content farther away in the network than the rest of active consumers ($\eta = 6.7 \cdot 10^{-8}$). Last,

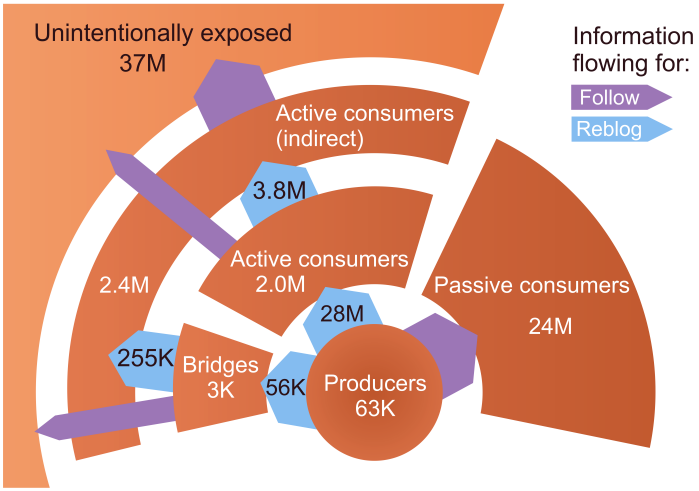


Figure 33: Tumblr - Diffusion of deviant content from the core of Producers to the rest of the network. Sectors represent disjoint user classes and arrows encode the information flow between them. Reblog arrows report the total volume of reblogs between two classes.

the audience of users who are potentially exposed in an unintentional way to deviant content includes almost 40M people. This figure should be considered as an upper bound on the number of people who actually have been exposed, as a follower of an active consumer might not see the pieces of deviant content for a number of reasons (e.g., inactivity, amount of content in the feed). That said, the pool of people who are potentially exposed is still very wide.

AQ4: *What is the perception of deviant content consumption from the perspective of individual nodes?*

Similar to real life, individuals in online social networks are most often aware of the activities of their direct social connections only but lack a global knowledge of the behavior of the rest of the population. In fact, the broad degree distribution of social networks may lead to the over-representation of rather rare nodal features when they observed in the

local context of an ego-network. This phenomenon has been observed in the form of the so-called *friendship paradox* (Fel91; HKL13), a statistical property of social networks for which on average people have fewer friends than their own friends. More recently the concept has been extended by the so-called *majority illusion* (LYW16), which states that in a social network with binary node attributes there might be a systematic local perception that the majority of people (50% or more) possess that attribute even when it is globally rare. As an illustrative example, in a network where people drinking alcohol are a small minority, the local perception of most nodes can be that the majority of people are drinkers just because drinkers happen to be connected with many more neighbors than the average. In our case study, active deviant content consumption is definitely a minority behavior compared to the 130M users in our sample. To estimate the presence of any skew in the local perception of deviant content consumption, we consider the nodes who are not producers and calculate the distribution of the proportion of their neighbors (in both the follow and reblog graphs) that either produce or reblog deviant material. The result is summarized in Figure 34. We observe that the follower network is nowhere close to exhibit the majority illusion phenomenon, with only the 10% of the population having 10% or more of their neighbors posting or reblogging deviant content. The effect increases sensibly when considering the reblog network, with 40% of the population locally observing more than 10% of their contacts reblogging deviant content and almost 10% having more than half of their neighbors doing it. This happens partly because the size of the reblog network is one order of magnitude smaller than the one of the follower network, as we consider reblogging activity for one month only. Still, this means that when looking at recent activity only, local perception biases are much stronger (although not predominant) in the community than what can be inferred from the static follow graph.

Although strongly biased perceptions are not predominant when counting the number of neighbors, a stronger bias emerges when looking at the *volume* of deviant content that is observed by a node from its neighbors. More than 71% of nodes reblogs less deviant content than the average of

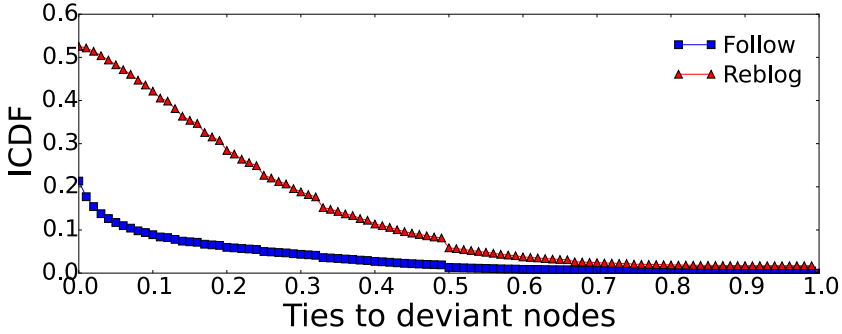


Figure 34: Proportion of nodes with at least a given ratio of outlinks landing on deviant nodes (inverse cumulative density function).

their friends (considering friends who posted or reblogged at least once in the time frame we consider). This effect, that derives directly from the strong correlation between degree and number of posts and reblogs, suggests that the local users' perception of other people's behavior is skewed towards an image of pervasive consumption of deviant content.

AQ5: *Is it possible to reduce the diffusion of deviant content with targeted interventions?*

Previous literature that investigated the properties of small-world networks indicates that information spreading or other phenomena of contagious nature can be drastically reduced by acting on a limited number of nodes in the graph (PSV05). Effectiveness of targeted interventions has been shown in a variety of domains, epidemics being the most prominent among them.

The intuition informed by previous work suggests that the wide diffusion of deviant content can be reduced by properly marking the posts produced by a small set of core nodes and showing them only to people who explicitly declared their interest for that specific topic. In a simplified experimental scenario, we measure the proportion of active consumers reached by adult content in a setting where all the posts from a set of core nodes C are erased. The question is how to select C and how big it needs

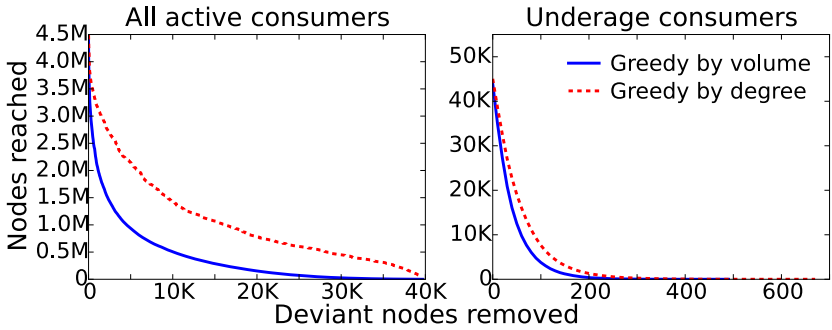


Figure 35: Shrinkage of content diffusion after deviant nodes removal, using two different strategies.

to be to uproot the diffusion process.

The optimal selection of nodes is a set cover problem (NP-complete), but we test two common approximated strategies to solve it: *i) greedy by volume*, an algorithm that ranks nodes by the number of blogs that are reached by the content they produce; and *ii) greedy by degree*, that takes into account the network structure only and ranks nodes by their in-degree in the reblog network. The effectiveness of the two approaches as $|C|$ increases in shown in Figure 35. Although using the indegree as proxy for the diffusion potential is not optimal, the removal of the $5{,}000$ highest indegree nodes curbs the diffusion by more than 50%. As expected, the strategy by volume is more effective (as it better approximates the optimal set cover), with a surprisingly sharp decay of the deviant content reach. The removal of the $5{,}000$ top nodes reduces the information spreading by nearly 80%, which increases to almost 100% when extending the block to $25{,}000$ nodes. Furthermore, using our sample of demographic information, we find that to limit the exposure of underage users would be sufficient to remove the 200 top nodes, as identified by any of the two selection strategies.

7.6.3 Demographics factors

The demographic composition of online adult content consumers has been measured by several sociological surveys (see Section 6.4), but none of them partitions the participants according to their type of consumption. Yet, we have shown that the categories of people exposed to online deviant content range from the active content producers to unintentional consumers. This calls for an investigation of the relationship between type of consumption and demographic characterization.

AQ6: *Is there a significant difference in the distribution of age and gender between members of the deviant network and people with different levels of exposure to deviant content?*

We report the distribution of age and gender of users with different levels of exposure to adult content, computed on the sample of 1.7M users who self-reported their demographic information. The average age in the sample is slightly higher than 26, and female are the majority (72%). To partly validate the user-provided information, we first compare them with third-party statistics. Our numbers are roughly compliant with several public reports that rely on orthogonal methods for assessing the age and gender of users (e.g., surveys and clickstream monitoring (Pin12; LaS12)). Those show that the Tumblr user base is the youngest among the most popular social networks and composed of women (65%) (Tay12). Also, we further validate the gender data by assessing that the 95% of users in the $Producer_2$ cluster focused on male homosexual content are indeed male. The overall age distribution of age by gender is shown in Figure 36: male tend to be older, originating a distribution with a fatter tail between age 35 and 55. Despite the spikes corresponding to birthdays in round decades (1970, 1980, and 1990), probably due to misreporting, the distribution still tends to be Gaussian, as expected.

We then measure differences in age[3] and gender distribution for the user classes of *producers*, *bridges*, *active consumers*, *passive consumers*, and

[3]The number of samples in each age distribution is high; therefore, as expected, all the differences between the average values are statistically significant ($p < 0.01$) under the Mann-Whitney test.

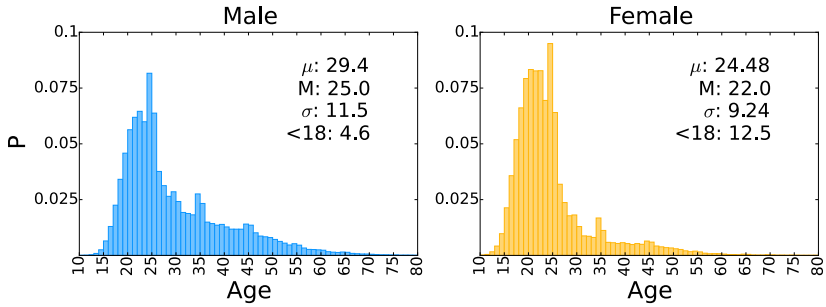


Figure 36: Age distribution of Tumblr users in our dataset. Mean μ, median M, standard deviation σ, and percentage of users under 18 years old are reported.

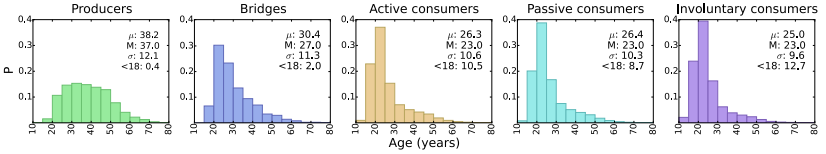


Figure 37: Age distribution of different groups of producers and consumers of adult content.

unintentionally exposed users (Figure 37). Producers are considerably older than the typical user, averaging around age 38 and with almost no underage users. Different from the overall distribution, they are mostly male (82%), in alignment with studies indicating that men are more involved in assiduous consumption of adult material. Bridge groups are fairly gender-balanced (with more female –68%– in the celebrity-oriented community) and include younger people (30 years old on average).
Consumers of deviant nodes who actively reblog or passively follow deviant blogs are covered by demographic data at 12%, proportion that drops to 4% among those who follow deviant nodes. In both classes, the age is quite representative of the overall Tumblr population in our sample (about 68% female). The same male-female proportion holds for people that are potentially exposed to deviant content in an unintentional way. This last class has the highest proportion of underage people (13%), which

reinforces the concern about young teens unwillingly seeing inappropriate content.
The fact that the gender distribution for active and passive consumers deviates only slightly from the overall gender distribution is in partial disagreement with previous studies on gender and sexual behavior (Hal06; KTLŠ14) which state that men are usually more exposed than women to adult material.
We conjecture that this might happen because of the tendency of female to have their peak of adult content consumption in a much younger age than men (as shown by (Fer03)), combined with the predominance of young female among Tumblr users. To verify it, we aim to answer one last question.

AQ7: *Does age have an effect on how different genders consume adult content?*
To find out, we measure the proportion of male and female actively exposed to deviant content (by reblogging), by age. We apply a min-max normalization to the obtained values so that scores towards 0 (1) represent the minimum (maximum) level of engagement. The curve for men shows an increasing trend that plateaus at its maximum in the range of age 35 to 55. In contrast, women, although less exposed than men at any age, have their peak in their 20s, much earlier than men. This observation supports previous findings (Fer03) and explains the distributions we observed.

7.6.4 Results in Flickr

We extended the present work with an equivalent study of deviant communities in Flickr. The study has been submitted to the journal Social Networks edited by Elsevier (CALS16b). We briefly report the results obtained in Flickr highlighting the differences.
While in Tumblr users share texts, images, video and audio, in Flickr the content of the social network is composed mainly by pictures and comments to them. To detect nodes dealing with pornography in Flickr we adopted a different strategy due to the fact that users usually do not access Flickr from a search engine but they access content by looking

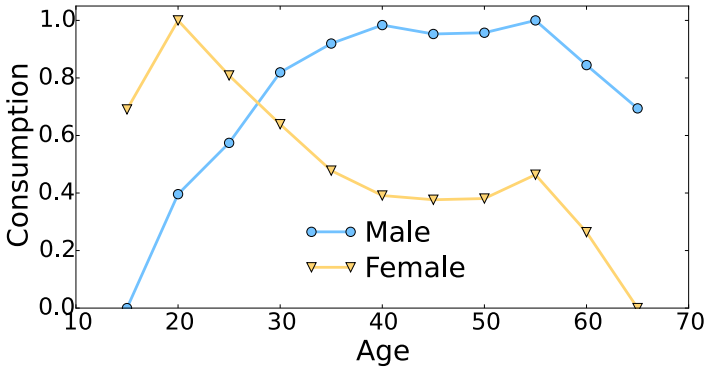


Figure 38: Ratio of male and female consuming adult content for different age bands in Tumblr (min-max normalized).

at their stream in the main page of the platform or through the internal search engine which integrates - however - stricter policies about showing adult content. Moreover, Flickr pictures are enriched with additional information that can be exploited to detect adult content. In particular the Flickr platform allows users to label their own photos as adult: around 423 thousand users uploaded pictures marked as adult. We found this labeling quite inaccurate, due to several reasons including semi-automatic batch photo tagging. Therefore, we exploited the vocabulary obtained through the *Deviant Graph Extraction* procedure to further refine the data collection mechanism by looking at photo tags. Each picture in Flickr is labeled with manual tags. On average each picture has 6 tags and each user publishes around 122 pictures.

We first slightly modified the adult vocabulary to remove misleading words in a photographic context (e.g., black&white) and we filtered photos labeled with at least one tag in the adult dictionary. We considered only users with at least two public adult photos identified as above in line with the query approach used for Tumblr where we marked a blog as adult only if it was reached through at least two unique adult queries. This procedure resulted in about 6.5 million photos by about 73 thousand deviant users.

In Flickr users share content not only in their profile but the platform enables the creation of groups whose members share images according to the topic of the group. To improve the recall of the data collection, we also identified those groups including at least one of the previously detected adult users and with a group title overlapping with adult vocabulary. All the users and photos in such groups were included in the adult cluster. Eventually, about 10M photos and 175 thousand *deviant* users were detected in Flickr.

Following the same approach used in analying Tumblr, we built two subgraphs of two networks induced by the *deviant nodes*: the follower/followee network and the favorite network. We used favorite links instead of the reblog links, as we did for Tumblr, since in Flickr the reblog action is not available. Basic network statistics on such subgraphs are reported in Table 26.

The deviant network is again a tiny portion of the whole graph, representing about 1.1% of all the nodes in the favorite graph in Flickr and a even smaller portion in the follow network (0.4%).

Compared to Tumblr the deviant community is bigger in Flickr, but with a comparable structure. We found ($Producers_1$) and ($Producers_2$) clusters with the same content characterization described for Tumblr and in addition to them a new cluster has been identified ($Producers_3$), whose users share mainly pictures representing transvestites or transsexual people. The same cluster is probably present in Tumblr but its size is not large enough to be distinguished by other producers. Producers clusters in Flickr are less than 58% of the deviant nodes with a large bridge cluster ($Bridge_1$) which is characterized by a content less explicit (soft porn, artistic nudity, hentai).

Clusters are detected with the same methodology used for Tumblr and they are shown in Figure 39.

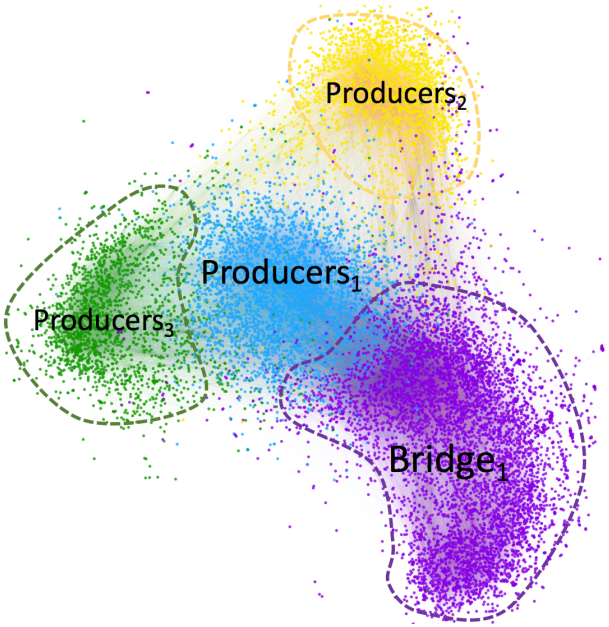


Figure 39: Bird-eye view of the deviant network in Flickr (favorite network), with colors denoting algorithmically-extracted communities.

From the bidimensional visualization of the network layout (Figure 39) it becomes apparent that again the *Producers*$_1$ and *Producers*$_2$ are two well-separated cores. *Producers*$_3$, is instead very close to *Producers*$_1$. The remaining community in Flickr is the largest cluster with a very elongated shape (showing the highest diameter), indicating a strong presence of soft content which is organized from a network point of view in a long cluster connected with the hard part only with one side. We named again this cluster *bridge community* as its main focus is often not on deviant content and it acts also as link towards the rest of the graph. Similarly we quantified the sizes of different classes in Flickr. Because of the absence of resharing actions in Flickr we calculated the size of *passive consumers* and *unintentionally exposed users* only. The results are summarized by the schema in Figure 40. The size of the producer clusters (90K) is almost 43% bigger than the case of Tumblr but still very small compared to the

Table 26: Flickr - Network statistics for the favorite (L) and follow (F) networks of the full graph sample (*All*), the deviant graph (*Deviant*), and the four communities that compose it (*Producers*$_{1,2,3}$ and *Soft*). All the statistics are about the giant weakly connected components and count only links whose both endpoints are in the considered node subset. $\langle k \rangle$=average degree, D=density, ρ=reciprocity, C=clustering, $\overline{spl}$=average shortest path length, d=diameter.

	$\lvert N \rvert$	$\lvert E \rvert$	$\langle k \rangle$	D	ρ	C	$\overline{spl}$	d
All L	15M	553M	37	$2{\cdot}10^{-6}$	0.06	-	-	-
All F	39M	566M	15	$4{\cdot}10^{-7}$	0.26	-	-	-
Deviant L	171K	13.4M	79	$5{\cdot}10^{-4}$	0.03	0.17	3.06	9
Deviant F	169K	37.9M	224	$1{\cdot}10^{-3}$	0.28	0.21	2.77	9
Bridge$_1$ L	66K	2.7M	47	$6{\cdot}10^{-4}$	0.05	0.17	3.05	13
Prod$_1$ L	53K	4.6M	99	$2{\cdot}10^{-3}$	0.03	0.18	2.83	13
Prod$_3$ L	20K	1.5M	94	$4{\cdot}10^{-3}$	0.03	0.23	2.53	13
Prod$_2$ L	16K	1.0M	83	$4{\cdot}10^{-3}$	0.04	0.28	2.52	14

whole network which is composed by 39M users. The size of the passive consumers is very consistent and comparable to the same class in Tumblr: around 20M users, mostly accessing deviant content by following the producers. The favorite action is quite limited: only 226K users exclusively like producers' content and 475K like and follow the producers at the same time. For Flickr the bridge cluster has a different role compared to Tumblr: resharing actions are not enabled but this community has an important role since its content is an entry point to access adult content by navigating the social network and the followers. In particular around 37% of the users in the bridge cluster follow the adult content producers and around 48% of them are in the group of users who like and follow the producers. Last, the audience of users who are potentially exposed in an unintentional way to deviant content includes almost 4.7M people: those are users who liked at least one picture from a another user who directly follows the deviant producers.

The pool of people who are exposed to adult content is quite significant even for Flickr even though is not comparable with the case of Tumblr and there are different reasons: the platform enable less sharing tools (no

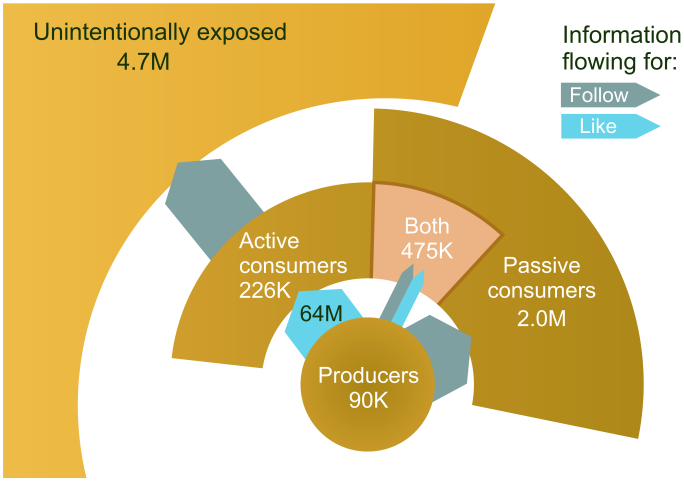


Figure 40: Flickr - Diffusion of deviant content from the core of Producers to the rest of the network. Sectors represent disjoint user classes and arrows encode the information flow between them.

reblog actions); the nature of the social network is more amateurish and the content is more personal. Tumblr blogs are more topic specific and the content uploaded is usually composed by multimedia material taken from other sources in the web which is specific for the adult sub-topic (e.g., 3d pornography, hentai, etc.), in Flickr instead each user shares with the communities pictures that in most of the cases are taken by himself/herself and they contain scenes of private life. The amateur nature of the content makes it of interest of family members, friends and close circles, preventing large diffusion.

Finally the demographic composition of Flickr is very different compared to Tumblr but similar properties are valid for this social network. In particular we considered 12.3M Flickr users who self-reported their demographic information. Flickr is more used by adult people and professional photographers with an age on average of 41 and it is more balanced in the gender of the users with 59% of male users. Producers are older than other categories (4 years more than average) and unintentional exposed

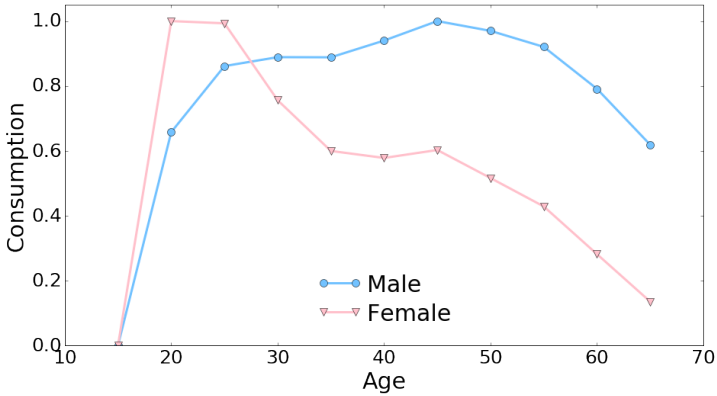


Figure 41: Ratio of male and female consuming adult content for different age bands in Flickr (min-max normalized).

users are the youngest as it is in Tumblr. Producers in Flickr, moreover, are for the 76% male confirming a higher presence of men in production of adult content. Last, we reproduced Figure 38 for Flickr (Figure 41) and we can see exactly the same trend with a female peak is consumption which is longer in Flickr up to 25, probably because the platform target a more adult audience.

7.7 Conclusion

This chapter aims to present a methodology to describe a community in an OSN looking at interactions among internal sub-group members and with the rest of the network, and studying the diffusion of the generated content and the role of the users in this process. In particular the focus of this chapter is on deviant communities: we intend to motivate researchers who study these groups online as well as offline to explore in more depth the interaction between the agents in such networks and the external social environment. Our contribution scratches only the surface of the exploration space that underlies the many types of networks - deviant and not - and the multitude of settings they are situated within. The study we

have presented is limited under many aspects, beginning from the focus on a single type of deviant behavior - adult material consumption - that is much more pervasive than others (e.g., anorexia) and, in that, has unique characteristics that likely cannot generalize to other deviant groups. In terms of methodology, alternative techniques (e.g., computer vision) could be used to identify adult content without a dedicated dictionary; those could possibly lead to describe the same phenomenon from a slightly different angle, for instance considering more exhaustively nodes that are not reached by search traffic.

We think it would be interesting to consider multiple deviant network types at different scales (e.g., content advocating violent behavior within the adult community). Also, we plan on analyzing the temporal dynamics of the deviant content spreading along social links.

Yet, we believe that our study has already important theoretical implications in revealing, for the first time on very large scale, that deviant communities can be deeply rooted into the relational fabric of a social network, and that the echo of their abnormal activity can reach a plenitude of ordinary users. Also, from a practical point of view, learning the effect that a minority group can have on a much larger audience is key to trigger mechanisms able to contain risky deviant phenomena by means of targeted interventions on few nodes, as we have shown.

We decided to include the study of deviant communities in this dissertation since it is a typology of group, often polarized, that is not frequently examined, mainly for lack of data, but that is salient to understand how people believe in digital contexts, stressing the difference with the *real world*, where people tend to expose less since their identity can not be hidden easily to the interlocutor.

Methodologically, this final chapter with all the previous ones, highlights the relevance of the study of content and interactions to explore common social behaviors, understanding social mechanisms that are the base of our nature: i.e., homophily, selection and access to information sources, creation of belief systems, content of interaction, content diffusion, social influence, membership. These are ingredients that must be deeply studied to undertake difficult challenges such as collective behavior prediction.

Conclusion

This dissertation is a study of *opinion polarization* on Online Social Networks (OSNs). The human need of creation of social bonds is one of the motivations for the popularity of Social Media (SM) nowadays. From a structural point of view, users' interaction network is often shaped in several dense communities, based on topics of discussion. Moreover such communities are often organized in polarized groups whose members share similar opinions. The identification and the tracking of the group's shared opinion is challenging because interactions are often characterized by short messages, harder to algorithmically "understand"; moreover, the comprehension of natural language opens interpretation problems and the users join and leave the conversation continuously, adding complexity to the process.

The first research question that we addressed is whether it possible to develop a methodology to detect in an automated way the topic of interaction and the opinions of the participants. We proposed a method to detect and track polarization by monitoring the evolution of users' interactions through an iterative classification of users and keywords: first, polarized users are identified, then polarized keywords are discovered by monitoring the activities of previously classified users. This method thus allows tracking users and topics over time. It is fast, flexible, and accurate, providing an improvement over network clustering baseline (k-means in in the word space of tweets) from 7% to 71% with different datasets of political tweets. Moreover, the possibility to couple together topic evolution and user polarization is a novel contribution of the method which

does not use any sentiment analysis and text processing methodology which have large limitations due to the uncertainty of natural languages. Note that our solution is almost completely unsupervised since just an initial seed of keywords per polarized group needs to be given as input to the algorithm. As a future work, the method could be fully automatized through auto-tuning of the parameters and self detection of the initial input seed based on the data.

The possibility to automatically detect *polarization* opens the discussion to the application areas of our and alternative approaches of opinion mining. The understanding of user opinions is important to consequently predict human behavior both in a collective way or individual by individual. OSNs are digital places to express desires, ideas, tastes, preferences, affiliations. Inferring this information from user activity is crucial for many disciplines, for instance in Marketing and in Sociology. In particular, one well studied application domain is the prediction of user political preferences. We focused on the possibilities and the limitations to use OSN data to predict the outcome of a political election and we proposed new strategies, showing a significant improvement over the baseline methods. On the other hand we confirmed the challenges of the task due to the bias of OSN users with respect to the general population (or to the voting population) and to the lack of data. Privacy limitations restrict the usable OSN data that we can collect to understand user behavior. To expand the knowledge about the users multiple OSNs should be integrated together, as well as other information sources (such as traditional polls, demographic data, historical data, analyses of events, etc.). The main issue is the matching problem in associating different users of multiple OSNs if they belong to the same physical individual. Even though we limit only to a single OSN much more information than *user polarization* can be extracted. The temporal dimension and the spacial one are very relevant to be taken into consideration if we want to extend the study of *polarization* more in detail. Along this research direction we selected the discussions about the Mediterranean refugee crisis over Twitter as a case study to show how to analyze *polarization* together with other variables provided by OSN data. We detailed the methodology to enrich

the raw stream of tweets with space information (user and mentioned locations), and sentiment (positive vs. negative) w.r.t. refugees. Our study shows differences in positive and negative sentiment in EU countries, in particular in UK, and by matching events, locations and perception, it underlines opinion dynamics and common prejudices regarding the refugees. For example we found that people who has a negative opinion about refugees prefer to call them migrants and link them with terrorism and Islamic fundamentalism. Still much more can be done in merging consistently alternative sources of information.

To further dig into polarized communities we focused on interactions both internally and among different communities. Sociological studies describe two situations that both might occur during interactions both in real and digital life: *echo chambers* and *controversy*. Both situations have been explored in our work. It is known that people trust sources someway consistent with their belief systems and recently the validity of the concept of collective intelligence in Social Media has been discussed because of the strong polarization which affects people, and then users, in accessing information. The fruition of knowledge is not "democratic", in the sense that in many situations people get informed only by sources in line with their believes, generating what it is usually referred as *cognitive closure*. We confirmed through a Facebook study the presence of this closed systems in which information and ideas are reinforced by internal transmission. These systems, generally called *echo chambers*, prevent users in changing opinion or at least in paying attention to a different point of view. In studying communities of Facebook users that retrieve information mostly from scientific pages and others that prefer sources based on conspiracy theories, we verified a *cognitive closure* effect in both groups, even if it was more evident in the second one. We then exposed both communities to a set of troll posts (false information) and we found that 77.92% of likes and 80.86% of comments were from users usually interacting with conspiracy stories, confirming the present of a strong echo chamber in this context.

The echo chamber encourages the interactions among members, but it is closed w.r.t. external communities. This does always not mean that interactions among different polarized communities are absent, while, on

the contrary, for some highly controversial topics (e.g., politics, religion, ethics) even though users prefer to get informed internally, they like to share their opinions and convictions with external users persuading them in joining their belief system or supporting, criticizing an event, a group, a party or a specific person. Controversial discussions then might be present among members of different polarized communities, even though these communities represent echo chambers, but cognitive closure of the participants is usually strong enough that the effect of this exchange of words reinforces user own opinion instead of reducing the polarization. *Controversy* then is a concept which is collateral to *polarization* and often high controversy implies high polarization. From a computer science perspective high controversy implies high interaction among members of different communities, which can be misinterpreted as closeness of the communities involved or even in the worse case as internal interactions among members of a unique community. For this reason using clustering methods to isolate polarized communities over the interaction graph has some limitations. Our approach in detecting polarization does not use network clustering methods but still it does not quantify the level of controversy among the communities. It is then interesting to go further with the research questions analyzing controversy and asking if it is possible to detect not only polarized users but also controversial discussions. We showed that it is feasible to detect controversy in Social Media by exploiting network motifs, i.e., local patterns of user interaction. The proposed approach allows for a language-independent and fine-grained analysis of user discussions and their evolution over time. We assessed the predictive power of motifs on a manually labeled Twitter dataset. The method does not use the content of the interaction to infer controversy and then again it is not subjected to the issues related to natural language processing. Our supervised model - exploiting motif patterns - achieved anyway 85% accuracy, with an improvement of 7% compared to structural, propagation-based and temporal network features. Additionally, thanks to the locality of motif patterns, we showed that it is possible to monitor the evolution of controversy in a conversation over time thus discovering sometimes changes in user opinion. Again a lot

of additional work can be done in understanding in which situations our approach works well and in which situations instead it is more effective to use content-based features.
The cognitive closure effect might be responsible to reduce content diffusion in the OSN. What about other social and psychological drivers that might reduce content spreading? In the final part of the thesis we focused on content dissemination. We targeted *deviant communities*, which are usually considered by their nature isolated, and, by analyzing content propagation in Tumblr and Flickr, we found instead that the produced content flows to the rest of the network mostly directly or through bridge-communities, reaching up to more than 450 times users. We also showed that a large fraction of the users can be inadvertently exposed to such content indirectly and we presented a demographic analysis of the producers and consumers networks. This final study show that content might spread even though the producer community is supposed to be isolated from the rest of the network. Our study is preliminary and, by targeting a specific topical community, we can not infer general properties. In any case we want to stress the attention on the need of analyzing more in detail additional drivers for popularity, virality and content spreading in general by looking also at particular social groups and their links with the whole OSN.
With this dissertation we hope to have shown the importance of studying user opinions and interactions in Social Media and to have contributed to the field by proposing novel methods to identify, analyze and track them automatically. We discussed solutions both under a computer science point of view and a sociological one, showing what can be extracted from Online Social Networks, how and for which purpose. We aim at setting new starting basis to study *opinion polarization* and *content spreading* in the context of Online Social Networks under a Computational Social Science perspective.

Glossary

CS = Computer Science A discipline based on the scientific and practical approach to computation and its applications.

CSS = Computational Social Science A new discipline based on the interdisciplinary investigation of the social universe on many scales, ranging from individual actors to the largest groupings, through the medium of computation (CR14).

Deviant behavior A conduct that is commonly considered inappropriate because it somehow violates society's norms or moral standards.

Deviant network A group of users interacting on a topic related to a deviant behavior.

Dunbar number A cognitive limit value to the amount of people with whom a person can maintain stable social relationships (≈ 150).

Echo chamber An *enclosed* system in which information, ideas, or beliefs are amplified or reinforced by internal transmission and repetition.

Ego network A focal node (*ego*) and the nodes to whom ego is directly connected (friends or alters) plus the ties, if any, among the alters.

Group or community A set of two or more people who interact with one another, share similar traits, and collectively have a sense of belonging.

ML = Machine Learning A sub-field of CS which evolved from the study of pattern recognition and computational learning theory in Artificial Intelligence (AI).

OSN = Online Social Network A platform to build social relations among people who share similar personal and career interests, activities, backgrounds or real-life connections (Bue16). Alternatively they are called Social Network Sites (SNS) or Social Media (SM).

Polarising sub-group A set of people sharing similar points of view about a specific discussed topic.

Social group A bond-based group characterized by personal social relations among members.

Social Network A structure made up of a set of actors, sets of dyadic ties, and interactions between individuals.

SS = Social Sciences A set of academic disciplines, concerning society and the relationships among individuals within it.

Topical group An identity-based group whose members share a common interest (topic).

Virtual world A computer-based simulated environment populated by users who simultaneously and independently explore the setting, participate in its activities and interact with others. An OSN is an example of virtual world as it is the web itself.

References

[ABC+10] Luca Maria Aiello, Alain Barrat, Ciro Cattuto, Giancarlo Ruffo, and Rossano Schifanella. Link creation and profile alignment in the aNobii social network. In *SocialCom*, 2010. 135

[ABS+12] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):9, 2012. 4, 11

[ACN14] Kelsey Allen, Giuseppe Carenini, and Raymond T Ng. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *EMNLP*, pages 1169–1180, 2014. 106

[ACPD13] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin Dunbar. Dynamics of personal social relationships in online social networks: A study on twitter. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 15–26, New York, NY, USA, 2013. ACM. 10

[ACPP12] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Analysis of ego network structure in online social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 31–40. IEEE, 2012. 11

[ADB95] Mike Allen, Dave D'Alessio, and Keri Brezgel. A meta-analysis summarizing the effects of pornography ii aggression after exposure. *Human communication research*, 22(2):258–283, 1995. 130

[AG05a] Lada Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *In LinkKDD 05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005. 83

[AG05b] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005. 23, 61, 105, 129

[AG12] Scott Atran and Jeremy Ginges. Religious and sacred imperatives in human conflict. *Science*, 336(6083):855–857, 2012. 97

[AH10] Sitaram Asur and Bernardo A 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. IEEE, 2010. 22, 38

[Aie15] Luca Maria Aiello. Group types in social media. In Georgios Paliouras, Symeon Papadopoulos, Dimitrios Vogiatzis, and Yiannis Kompatsiaris, editors, *User Community Discovery*, Human-Computer Interaction Series, pages 97–134. Springer International Publishing, 2015. 5, 12, 129

[All12] James Allan. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media, 2012. 23

[Amb13] Giulio Ambrosetti. I forconi: il senato ha approvato una legge per i parlamentari in crisi. chi non verr rieletto, oltre alla buonuscita, si beccher altri soldi. sar vero? Website, 8 2013. last checked: 19.01.2014. 83, 97

[AP46] GORDON W. ALLPORT and LEO POSTMAN. An analysis of rumor. *Public Opinion Quarterly*, 10(4):501–517, 1946. 81

[AQC14] Jisun An, Daniele Quercia, and Jon Crowcroft. Partisan sharing: facebook evidence and societal consequences. In *COSN*, pages 13–24, 2014. 105

[AR98] Michael Ayers and Lynne Reder. A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review*, 5(1):1–21, March 1998. 82

[Att05] Feona Attwood. What do people do with porn? qualitative research into the comsumption, use, and experience of pornography and other sexually explicit media. *Sexuality and culture*, 9(2), 2005. 126

[BAA05] Orkut Buyukkokten, Eytan Adar, and Lada Adamic. A social network caught in the web. *First Monday*, 2005. 9

[Bau97] M. Bauer. *Resistance to New Technology: Nuclear Power, Information Technology and Biotechnology*. Cambridge University Press, 1997. 82

[BAX10] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. Investigating homophily in online social networks. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 533–536. IEEE, 2010. 4

[BBEM11] Victor Bekkers, Henri Beunders, Arthur Edwards, and Rebecca Moody. New media, micromobilization, and political agenda setting: Crossover effects in political mobilization and media usage. *The Information Society*, 27(4):209–219, July 2011. 82

[BBM13] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Cascade-based community detection. In *WSDM*. ACM, 2013. 12, 129

[BBR+12] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 33–42. ACM, 2012. 9

[BC15] Ruggero Bellio and Mauro Coletto. Simple outlier labelling based on quantile regression, with application to the steelmaking process. *Applied Stochastic Models in Business and Industry*, 2015. 19

[BCD+14a] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Sensing information-based communities in the age of misinformation. In *ECCS 2014, Lucca, Italy*, 2014. xix, 15, 17, 79

[BCD+14b] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, and Walter Quattrociocchi. Misinformation in the loop: the emergence of narratives in osn. *ITAIS 2014, Genova, Italy*, 2014. xix, 15, 17, 79

[BCD+15] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2), 2015. xix, 13, 15, 17, 23, 79, 86

[BCDV+14] Alessandro Bessi, Guido Caldarelli, Michela Del Vicario, Antonio Scala, and Walter Quattrociocchi. Social determinants of content selection in the age of (mis)information. In LucaMaria Aiello and Daniel McFarland, editors, *Social Informatics*, volume 8851 of *Lecture Notes in Computer Science*, pages 259–268. Springer International Publishing, 2014. 81, 101

[BDGL08] Ilaria Bordino, Debora Donato, Aristides Gionis, and Stefano Leonardi. Mining large networks with subgraph counting. In *2008 Eighth IEEE International Conference on Data Mining*, pages 737–742. IEEE, 2008. 114

[BF10] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010. 107

[BFJ+12] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, September 2012. 83

[BGL16] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016. 106

[BGLL08a] Vincent D. Blondel, Jean L. Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of community hierarchies in large networks, March 2008. 87, 93

[BGLL08b] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. 136

[BHKL06] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006. 11

[BHR07] Louise Barrett, Peter Henzi, and Drew Rendall. Social brains, simple minds: does social complexity really require cognitive complexity? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1480):561–575, 2007. 3

[BHRG+11] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz, Francisco Sanz, Fermín Serrano, Cristina Viñas, Alfonso Tarancón, Yamir Moreno, and Matjaz Perc. Structural and dynamical patterns on online social networks: The spanish may 15th movement as a case study. *PLoS One*, 6(8):e23883, Aug 2011. 83

[BL07] David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007. 12

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011. 22, 38

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 12

[BnKVR03] E. Ben-naim, P. L. Krapivsky, F. Vazquez, and S. Redner. Unity and discord in opinion dynamics. *Physica A*, 2003. 83

[BP12] Natalie Boero and Cheri Jo Pascoe. Pro-anorexia communities and online interaction: Bringing the pro-ana body online. *Body & Society*, 18(2):27–57, 2012. 129

[BRMA12] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012. xviii, 100

[BS10] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM, 2010. 40

[BS11] Adam Bermingham and Alan F Smeaton. On using twitter to monitor political sentiment and predict election results. 2011. 39, 40

[BSAS+12] S. Buckingham Shum, K. Aberer, A. Schmidt, S. Bishop, P. Lukowicz, S. Anderson, Y. Charalabidis, J. Domingue, S. Freitas, I. Dunwell, B. Edmonds, F. Grey, M. Haklay, M. Jelasity, A. Karpitenko, J. Kohlhammer, J. Lewis, J. Pitt, R. Sumner, and D. Helbing. Towards a global participatory platform. *The European Physical Journal Special Topics*, 214(1):109–152, 2012. 82

[BSZ+ar] Alessandro Bessi, Antonio Scala, Qian Zhang, Luca Rossi, and Walter Quattrociocchi. The economy of attention in the age of (mis)information. *Journal of Trust Management*, to appear. 83, 97

[Bue16] Ricardo Buettner. Getting a job via career-oriented social networking sites: The weakness of ties. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 2156–2165. IEEE, 2016. 160

[Buz05] Timothy Buzzell. Demographic characteristics of persons using pornography in three technological contexts. *Sexuality & Culture*, 9(1):28–48, 2005. 130

[Byf11] J. Byford. *Conspiracy Theories: A Critical Introduction*. Palgrave Macmillan, 2011. 82, 98

[CAB+12] Peter Cogan, Matthew Andrews, Milan Bradonjic, W Sean Kennedy, Alessandra Sala, and Gabriel Tucci. Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 25–31. ACM, 2012. 105

[CAD+14] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee. 83

[CALS16a] Mauro Coletto, Luca Maria Aiello, Claudio Lucchese, and Fabrizio Silvestri. On the behaviour of deviant communities in online social networks. In *ICWSM 2016, Köln, Germany*, 2016. xix, 13, 15, 18, 125

[CALS16b] Mauro Coletto, Luca Maria Aiello, Claudio Lucchese, and Fabrizio Silvestri. On the behaviour of deviant communities in online social networks. *SUBMITTED TO SOCIAL NETWORKS (ELSEVIER)*, 2016. xix, 15, 18, 125, 147

[CB13] Zoey Chen and Jonah Berger. When, why, and how controversy causes conversation. *Journal of Consumer Research*, 40(3):580–593, 2013. 106

[CCP+14] Guido Caldarelli, Alessandro Chessa, Fabio Pammolli, Gabriele Pompa, Michelangelo Puliga, Massimo Riccaboni, and Gianni Riotta. A multi-level geographical study of italian political elections from twitter data. *PloS one*, 9(5):e95809, 2014. 39

[CDCS10] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM, 2010. 23

[CDF+13] Michael D Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial

characteristics of a social movement communication network. *PloS one*, 8(3), 2013. 61

[Cen10] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, September 2010. 82, 83

[CF08] Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258, 2008. 130

[CFL09] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591+, June 2009. 83, 84, 94

[CGB+12] Rosaria Conte, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, J-P Nadal, Anxo Sanchez, et al. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346, 2012. 6, 7

[CGGL17] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. A motif-based approach for identifying controversy. In *WSDM 2017, Cambridge, UK*, 2017. xix, 18, 100

[Cir14a] Facebook, trolls, and italian politics, March 2014. 96

[Cir14b] Study explains why your stupid facebook friends are so gullible, March 2014. 96

[CJM10] Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. Identifying controversial issues and their sub-topics in news articles. In *Intelligence and Security Informatics*, pages 140–153. Springer, 2010. 105

[CL16] Mauro Coletto and Claudio Lucchese. Social-spatio-temporal analysis of topical and polarised communities in online social networks. In *Encyclopedia of Social Network Analysis and Mining*, pages 0–0. Springer, 2016. 2

[CLCY13] An-Sing Chen, Mark Leung, Chih-Hao Chen, and Shu Ching Yang. Exposure to internet pornography among taiwanese adolescents. *Social Behavior and Personality: an international journal*, 41(1):157–164, 2013. 130

[CLM+16] Mauro Coletto, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Andrea Esuli, Chiara Renso, and Raffaele Perego. Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis. In *SNAST Workshop on Social Network Analysis Surveillance Technologies, ASONAM 2016, San Francisco, California*, 2016. xix, 15, 16, 56

[CLO+15] Mauro Coletto, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Alessandro Chessa, and Michelangelo Puliga. Electoral predictions with twitter: a joint machine learning and complex network approach applied to an italian case study. In *ICCSS 2015, Helsinki, Finland*, 2015. xix, 15, 37

[CLOP15] Mauro Coletto, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Electoral predictions with twitter: a machine-learning approach. In *IIR 2015, Cagliari, Italy*, 2015. xix, 15, 16, 23, 33, 37, 38, 61

[CLOP16] Mauro Coletto, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Polarized user and topic tracking in twitter. In *SIGIR 2016, Pisa, Italy*, 2016. xix, 15, 20, 37, 56, 62, 64, 65, 105

[CM15] Marshall Clinard and Robert Meier. *Sociology of deviant behavior*. Wadsworth Cengage Learning, 2015. xx, 126

[CMP11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011. 106

[CNM04] Aaron Clauset, M. E. J. Newman, , and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, pages 1– 6, 2004. 87, 93

[CR10] Claudio Cioffi-Revilla. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):259–271, 2010. 6

[CR14] Claudio Cioffi-Revilla. *Introduction to Computational Social Science: Principles and Applications*. Springer Publishing Company, Incorporated, 2014. 160

[CRF+11a] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *ICWSM*, 2011. 12, 129

[CRF+11b] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *ICWSM*, 2011. 23, 24, 102, 105

[Cri14] Nello Cristianini. On the current paradigm in artificial intelligence. *AI Communications*, 27(1):37–43, 2014. 5

[DC15] Munmun De Choudhury. Anorexia on tumblr: A characterization study. In *Digital Health*. ACM, 2015. 126, 129

[DDLMM13] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3, 2013. 106

[DHA13] Shiri Dori-Hacohen and James Allan. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1845–1848. ACM, 2013. 104

[DMBR13] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11):e79449, 2013. 23, 39, 40, 42, 46

[DNAW01] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3:87–98, 2001. 84, 94

[Dun92] Robin IM Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992. 11

[Dun93] Robin IM Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(04):681–694, 1993. 11

[EK10] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010. 129

[EPDRH96] Seymour Epstein, Rosemary Pacini, Veronika Denes-Raj, and Harriet Heier. Individual differences in intuitiveexperiential and analyticalrational thinking styles. *Journal of Personality and Social Psychology*, 71(2):390–405, 1996. 84, 91

[ESL07] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. The benefits of facebook friends: social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, July 2007. 84

[Fac13] Facebook. Using the graph api. Website, 8 2013. last checked: 19.01.2014. 86

[FAcC] Adrien Friggeri, Lada Adamic, Dean ckles, and Justin Cheng. Rumor Cascades. *AAAI Conference on Weblogs and Social Media (ICWSM)*. 83

[Fan10] Daniele Fanelli. Do pressures to publish increase scientists' bias? an empirical support from us states data. *PloS one*, 5(4):e10271, 2010. 40

[FCVH] G.A. Fine, V. Campion-Vincent, and C. Heath. *Rumor Mills: The Social Impact of Rumor and Legend*. Social problems and social issues. Transaction Publishers. 82, 98

[Fel91] Scott L Feld. Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477, 1991. 142

[Fer03] Marnie Ferree. Women and the web: Cybersex activity and implications. *Sexual and Relationship Therapy*, 18(3):385–393, 2003. 147

[FKSW11] Albert Feller, Matthias Kuhnert, Timm Oliver Sprenger, and Isabell M Welpe. Divided they tweet: The network structure of political microbloggers and discussion topics. In *ICWSM*, 2011. 129

[FNL11] Steven J. Frenda, Rebecca M. Nichols, and Elizabeth F. Loftus. Current Issues and Advances in Misinformation Research. *Current Directions in Psychological Science*, 20:20–23, 2011. 83

[For10] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010. xviii, 100, 136

[GA11] Daniel Gayo-Avello. Don't turn social media into another 'literary digest' poll. *Communications of the ACM*, 54(10):121–128, 2011. 40

[GAEJ13] Przemyslaw A. Grabowicz, Luca Maria Aiello, Victor M. Eguiluz, and Alejandro Jaimes. Distinguishing topical and social groups based on common identity and bond theory. In *WSDM*. ACM, 2013. 129

[GAMM11] Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *ICWSM*, 2011. 23, 40

[GDFMGM16] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy in social media. In *ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 33–42, 2016. 24, 61, 62, 102, 103, 105, 108, 121

[Gig12] Fabio Giglietto. If likes were votes: An empirical study on the 2011 italian administrative elections. 2012. 39

[GMJCK13] Pedro Henrique Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. A measure of polarization on social media networks based on community boundaries. In *ICWSM*, 2013. 21, 105, 129

[GN02] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002. 4

[GPV11] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, 2011. 11

[Gra73] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973. 3, 10

[GRP08] Jeff Gavin, Karen Rodham, and Helen Poyer. The presentation of "pro-anorexia" in online group interactions. *Qualitative Health Research*, 18(3):325–333, 2008. 129

[GW13] R. Kelly Garrett and Brian E. Weeks. The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 1047–1058, New York, NY, USA, 2013. ACM. 82

[Hal06] Gert Martin Hald. Gender differences in pornography consumption among young heterosexual danish adults. *Archives of sexual behavior*, 35(5):577–585, 2006. 130, 147

[HB11] M.A. Hogg and D.L. Blaylock. *Extremism and the Psychology of Uncertainty*. Blackwell/Claremont Applied Social Psychology Series. Wiley, 2011. 82, 98

[HIJW10] Stephen M Haas, Meghan E Irr, Nancy A Jennings, and Lisa M Wagner. Online negative enabling support groups. *New Media & Society*, 2010. 126, 129

[HKL13] Nathan O Hodas, Farshad Kooti, and Kristina Lerman. Friendship paradox redux: Your friends are more interesting than you. In *ICWSM*, 2013. 142

[HMKW14] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, Ann Arbor, MI, June 2014. 83

[HML13] Gert Martin Hald, Neil N Malamuth, and Theis Lange. Pornography and sexist attitudes among heterosexuals. *Journal of Communication*, 63(4):638–660, 2013. 130

[How13] Lee Howell. Digital wildfires in a hyperconnected world. In *Report 2013*. World Economic Forum, 2013. 82

[HŠ15] Gert Martin Hald and Aleksandar Štulhofer. What types of pornography do people use and do they cluster? assessing types and categories of pornography consumption in a large-scale online sample. *The Journal of Sex Research*, pages 1–11, 2015. 130

[HSB+14] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014. 61

[Ise86] Daniel J Isenberg. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology*, 50(6):1141, 1986. xix, 21

[JJS12] Andreas Jungherr, Pascal Jürgens, and Harald Schoen. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im predicting elections with twitter: What 140 characters reveal about political sentiment. *Social Science Computer Review*, 30(2):229–234, 2012. 40

[Joi08] Adam N. Joinson. Looking at, looking up or keeping up with people?: Motives and use of facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1027–1036, New York, NY, USA, 2008. ACM. 84

[KG14] Simone Kühn and Jürgen Gallinat. Brain structure and functional connectivity associated with pornography consumption: the brain on porn. *JAMA psychiatry*, 71(7):827–834, 2014. 130

[KGP00] A. Koriat, M. Goldsmith, and A. Pansky. Toward a psychology of memory accuracy. *Annu Rev Psychol*, 51:481–537, 2000. 82

[KKNG12] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. Geographic dissection of the Twitter network. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012. 60

[Kle13] Jon Kleinberg. Analysis of large-scale social and information networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371, 2013. 83

[KLNN+05] Ravi Kumar, David Liben-Nowell, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Theoretical analysis of geographic routing in social networks. 2005. 11

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010. xviii, 9, 100

[KNT10] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer, 2010. 10

[KPV+14] Andreas Kanavos, Isidoros Perikos, Pantelis Vikatos, Ioannis Hatzilygeroudis, Christos Makris, and Athanasios Tsakalidis. Conversation emotional modeling in social networks. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, pages 478–484. IEEE, 2014. 106

[KQJ+00] James H Kuklinski, Paul J Quirk, Jennifer Jerit, David Schwieder, and Robert F Rich. Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3):790–816, 2000. 81

[KSTA15] Olga Kolchyna, Tharsis TP Souza, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*, 2015. 61

[KTLŠ14] Ingela Lundin Kvalem, Bente Træen, Bo Lewin, and Aleksandar Štulhofer. Self-perceived effects of internet pornography use, genital appearance satisfaction, and sexual self-esteem among young scandinavian adults. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8(4), 2014. 147

[Kuh62] Thomas S Kuhn. The structure of scientific revolutions, 1962. 5

[LAH07] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007. 16, 38

[LaS12] Ryan LaSala. The social makeup of social media. *Compete.com Tech Blog*, 2012. 145

[LC11] Lung-Hao Lee and Hsin-Hsi Chen. Collaborative cyberporn filtering with collective intelligence. In *SIGIR*. ACM, 2011. 131

[LCFT04] Pamela J Ludford, Dan Cosley, Dan Frankowski, and Loren Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 631–638. ACM, 2004. 11

[LCN15] Haokai Lu, James Caverlee, and Wei Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 213–222. ACM, 2015. 24, 102, 105

[LCOM13] Stephan Lewandowsky, John Cook, Klaus Oberauer, and Michael Marriott. Recursive fury: Conspiracist ideation in the blogosphere in response to research on conspiracist ideation. *Frontiers in Psychology*, 4, 2013. 97

[LDBC10] Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector-tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*, pages 599–602. Springer, 2010. 22, 38

[Lee97] Roger Leenders. Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. *Evolution of social networks*, 1, 1997. 11

[LGK12] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, January 2012. 83

[LH08] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008. 11

[LKF05] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *ACM SIGKDD*. ACM, 2005. 138

[LNK07] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007. xviii, 100

[LPA+09] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009. 38

[LSM11] Jimmy Lin, Rion Snow, and William Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *ACM SIGKDD*, pages 422–429. ACM, 2011. 22, 24

[LYW16] Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu. The "majority illusion" in social networks. *PLoS ONE*, 11(2):1–13, 02 2016. 142

[Man96] Klaus Manhart. 19 artificial intelligence modelling: Data driven and theory driven approaches. *Social Science Microsimulation*, 1996. 5

[Mas43] Abraham Harold Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943. 3

[MB08] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008. 12

[MBAG+14a] David Martin-Borregon, Luca M Aiello, Przemyslaw Grabowicz, Alejandro Jaimes, and Ricardo Baeza-Yates. Characterization of online groups along space, time, and social dimensions. *EPJ Data Science*, 3(1):8, 2014. 11

[MBAG+14b] David Martin-Borregon, Luca Maria Aiello, Przemyslaw Grabowicz, Alejandro Jaimes, and Ricardo Baeza-Yates. Characterization of online groups along space, time, and social dimensions. *EPJ Data Science*, 3(1):8, 2014. 129

[MBG+13] Delia Mocanu, Andrea Baronchelli, Bruno Gonçalves, Nicola Perra, Qian Zhang, and Alessandro Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PLOS ONE*, 8(4):e61981, 2013. 83

[MBLB15] Alfredo J Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015. 105

[MC84] Peter V Marsden and Karen E Campbell. Measuring tie strength. *Social forces*, 63(2):482–501, 1984. 10

[MCST+12] Gabriel Magno, Giovanni Comarela, Diego Saez-Trumper, Meeyoung Cha, and Virgilio Almeida. New kid on the block: Exploring the google+ social graph. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 159–170. ACM, 2012. 9

[MFW03] Kimberly J Mitchell, David Finkelhor, and Janis Wolak. The exposure of youth to unwanted sexual material on the internet a national survey of risk, impact, and prevention. *Youth & Society*, 34(3):330–358, 2003. 130

[Mil67] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967. 3

[MK10] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proc. of the 2010 ACM SIGMOD*, pages 1155–1158. ACM, 2010. 22, 23

[MM13] Karissa McKelvey and Filippo Menczer. Truthy: Enabling the study of online social networks. In *Proc. CSCW '13*, 2013. 82

[MMG+07] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007. 9

[MMGA11] Panagiotis Takis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (not) to predict elections. In *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*, pages 165–171. IEEE, 2011. 40

[MR02] Michelle L. Meade and Henry L. Roediger. Explorations in the social contagion of memory. *Memory & Cognition*, 30(7):995–1009, 2002. 82

[MRZ+14] Delia Mocanu, Luca Rossi, Qian Zhang, Màrton Karsai, and Walter Quattrociocchi. Collective attention in the age of (mis)information. *CoRR*, abs/1403.3344, 2014. 83, 97

[MS72] M. E. Mccombs and D. L. Shaw. The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly*, 36(2):176–187, 1972. 81

[MS00] Chris Mann and Fiona Stewart. *Internet Communication and Qualitative Research: A Handbook for Researching Online (New Technologies for Social Research series)*. Sage Publications Ltd, September 2000. 82

[MSEB10] Elizabeth M Morgan, Chareen Snelson, and Patt Elison-Bowers. Image and video disclosure of substance use on social media websites. *Computers in Human Behavior*, 26(6):1405–1411, 2010. 126, 129

[MSL08] Paul R Messinger, Eleni Stroulia, and Kelly Lyons. A typology of virtual worlds: Historical overview and future directions. *Journal For Virtual Worlds Research*, 1(1), 2008. 7

[MSLC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001. 11

[MSOI+02] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. 106, 114

[MZDC14] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. Controversy and sentiment in online news. *Symposium on Computation + Journalism*, 2014. 23, 102, 105

[NG04] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004. 5

[NGP08] Radu Andrei Negoescu and Daniel Gatica-Perez. Analyzing flickr groups. In *CIVR*, New York, NY, USA, 2008. ACM. 129

[NTO+16] Ryosuke Nishi, Taro Takaguchi, Keigo Oka, Takanori Maehara, Masashi Toyoda, Ken-ichi Kawarabayashi, and Naoki Masuda. Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining*, 6(1):1–13, 2016. 105

[OAG+11] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic constraints on social network groups. *PLoS one*, 6(4):e16939, 2011. 63

[OBRS10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010. 39

[OOL14] Hyun Jung Oh, Elif Ozkaya, and Robert LaRose. How does online social networking enhance life satisfaction? the relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. *Computers in Human Behavior*, 30:69–78, 2014. 8

[ORT10] Jukka-Pekka Onnela and Felix Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences*, 107(43):18375–18380, October 2010. 83

[PEJ+09] M. Paolucci, T. Eymann, W. Jager, J. Sabater-Mir, R. Conte, S. Marmo, S. Picascia, W. Quattrociocchi, T. Balke, S. Koenig, T. Broekhuizen, D. Trampe, M. Tuk, I. Brito, I. Pinyol, and D. Villatoro. *Social Knowledge for e-Governance: Theory and Technology of Reputation*. Roma: ISTC-CNR, 2009. 82, 83

[Pin12] Pingdom. Report: Social network demographics in 2012. *Pingdom.com Tech Blog*, 2012. 145

[PJL13] Marco Cinnirella Patrick J. Leman. Beliefs in conspiracy theories and the need for cognitive closure, 2013. 84, 91

[PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008. 61

[PML94] Deborah A Prentice, Dale T Miller, and Jenifer R Lightdale. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Key Readings in Social Psychology*, page 83, 1994. 11

[PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010. 61

[PSV05] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemics and immunization in scale-free networks. *Handbook of graphs and networks: from the genome to the internet*, pages 111–130, 2005. 143

[QCL11] W. Quattrociocchi, R. Conte, and E. Lodi. Opinions manipulation: Media, power and gossip. *Advances in Complex Systems*, 14(4):567–586, 2011. 82, 83

[QCS14] Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. Opinion dynamics on interacting networks: media competition and social influence. *Scientific Reports*, 4, May 2014. 82, 83

[QPC09] Walter Quattrociocchi, Mario Paolucci, and Rosaria Conte. On the effects of informational cheating on social evaluations: image and reputation through gossip. *IJKL*, 5(5/6):457–471, 2009. 82, 83

[RB15] Patrizia Romito and Lucia Beltramini. Factors associated with exposure to violent or degrading pornography among high school students. *The Journal of School Nursing*, 31(4):280–90, 2015. 130

[RCC15] Manuela Ritondale, Guido Caldarelli, and Mauro Coletto. Application of network analysis to the trade routes of antiquities passing through the pontine islands. In *CAA 2015, Siena, Italy*, 2015. 19

[RCM+11] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. In *WWW*, 2011. 13

[RM14] Andrew Rojecki and Sharon Meraz. Rumors and factitious informational blends: The role of the web in speculative politics. *New Media and Society*, 2014. 81

[RPNB11] Juliana de Souza Ramos, André de Faria Pereira Neto, and Marcos Bagrichevsky. Pro-anorexia cultural identity: characteristics of a lifestyle in a virtual community. *Interface (Botucatu)*, 15(37):447–460, Jun 2011. 129

[RTU13] Daniel M Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. In *ICWSM*, 2013. 129

[SB12] Erik Tjong Kim Sang and Johan Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. 39, 40, 44

[SBC+10] Rossano Schifanella, Alain Barrat, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Folks in folksonomies: Social link prediction from shared metadata. In *WSDM*. ACM, 2010. 138

[SC15] Celeste Stefania and Mauro Coletto. La sentiment analysis per i musei 2.0. un approccio bottom-up per la conoscenza del pubblico, 2015. 19

[sim14a] simplyhumans. Dieta peronalizzata - benefic effect of lemons. Website, 7 2014. last checked: 31.07.2014. 83

[sim14b] simplyhumans. Simply humans - benefic effect of lemons. Website, 7 2014. last checked: 31.07.2014. 83

[sim14c] simplyhumans. Simply humans - benefic effect of lemons. Website, 7 2014. last checked: 31.07.2014. 83

[SMML10] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. Distance matters: Geo-social metrics for online social networks. In *Conference on Online Social Networks*, WOSN'10, 2010. 60

[SPA+12] Marko Skoric, Nathaniel Poor, Palakorn Achananuparp, Ee-Peng Lim, and Jing Jiang. Tweets and votes: A study of the 2011 singapore general election. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2583–2591. IEEE, 2012. 40

[SR08] Aaron Whitman Smith and Harrison Rainie. *The Internet and the 2008 election*. Pew Internet & American Life Project, 2008. 40

[SSC16] Benjamin Shulman, Amit Sharma, and Dan Cosley. Predictability of popularity: Gaps between prediction and understanding. *ICWSM*, 2016. 119

[Sun02] Cass R Sunstein. The law of group polarization. *Journal of political philosophy*, 10(2):175–195, 2002. xix, 21

[Sur05] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*. Abacus, 2005. 82

[SV09] Cass R. Sunstein and Adrian Vermeule. Conspiracy theories: Causes and cures*. *Journal of Political Philosophy*, 17(2):202–227, June 2009. 82

[SWF08] Chiara Sabina, Janis Wolak, and David Finkelhor. The nature and dynamics of internet pornography exposure for youth. *CyberPsychology & Behavior*, 11(6), 2008. 128, 130

[SZV13] Michael Schuhmacher, Cäcilia Zirn, and Johanna Völker. Exploring youporn categories, tags, and nicknames for pleasant recommendations. In *Workshop on Search and Exploration of X-Rated Information*. ACM, 2013. 130

[Taj82] Henri Tajfel. Social psychology of intergroup relations. *Annual review of psychology*, 33(1):1–39, 1982. 4

[Tay12] Chris Taylor. Women win facebook, twitter, zynga; men get linkedin, reddit. *Mashable.com*, Jul 2012. 145

[TBT+11] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011. 61

[TESU13] Gareth Tyson, Yehia Elkhatib, Nishanth Sastry, and Steve Uhlig. Demystifying porn 2.0: A look into a major adult video streaming website. In *IMC*. ACM, 2013. 130

[TESU15] Gareth Tyson, Yehia Elkhatib, Nishanth Sastry, and Steve Uhlig. Are People Really Social in Porn 2.0? In *ICWSM*, May 2015. 126, 129

[TGW12] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of Twitter networks. *Social Networks*, 34(1):73 – 81, 2012. 60

[TSSW10] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010. 23, 33, 39, 40, 42, 45, 53

[Tur81] John C Turner. Towards a cognitive redefinition of the social group. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 1981. 4

[UBMK12] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 2012. 83

[UKBM11] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011. 9

[VMCG09] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09, pages 37–42, New York, NY, USA, 2009. ACM. 84, 93

[Wat17] Duncan J Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1:0015, 2017. 6

[WBBP10] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432, 2010. 82

[WBS+09] Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. In *Algorithmic Aspects in Information and Management*, pages 301–314. Springer, 2009. 12

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. 2, 3

[WG08] Christine Williams and Girish Gulati. What is a social network worth? facebook and vote share in the 2008 presidential primaries. American Political Science Association, 2008. 39

[WJSS99] Frederick Walls, Hubert Jin, Sreenivasa Sista, and Richard Schwartz. Topic detection in broadcast news. In *Proceedings of the DARPA broadcast news workshop*, pages 193–198, 1999. 23

[WK94] Donna M. Webster and Arie W. Kruglanski. Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67(6):1049–1062, 1994. 84, 91

[WL11] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011. 66

[WMF07] Janis Wolak, Kimberly Mitchell, and David Finkelhor. Unwanted and wanted exposure to online pornography in a national sample of youth internet users. *Pediatrics*, 119(2):247–257, 2007. 130

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998. 3

[WSPZ12] Christo Wilson, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web (TWEB)*, 6(4):17, 2012. 9, 10

[XC08] Jennifer Xu and Hsinchun Chen. The topology of dark networks. *Communications of the ACM*, 51(10):58–65, 2008. 130

[YKSG14] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916. ACM, 2014. 24

[YM05] Michele L Ybarra and Kimberly J Mitchell. Exposure to internet pornography among children and adolescents: A national survey. *CyberPsychology & Behavior*, 8(5):473–486, 2005. 130

[ZCL+10] Bi Zhu, Chuansheng Chen, Elizabeth F. Loftus, Chongde Lin, Qinghua He, Chunhui Chen, He Li, Robert K. Moyzis, Jared Lessard, and Qi Dong. Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. *Personality and Individual Differences*, 48(8):889 – 894, 2010. 83

[ZGDNM16] Muhammad Bilal Zafar, Krishna P Gummadi, and Cristian Danescu-Niculescu-Mizil. Message impartiality in social media discussions. In *Tenth International AAAI Conference on Web and Social Media*, 2016. 106

[ZGWS14] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. Inferring international and internal migration patterns from twitter data. In *WWW Conference*, WWW'14 Companion, 2014. 61