ANALYZING SOCIAL BIG DATA TO STUDY ONLINE DISCOURSE AND ITS MANIPULATION

Onur Varol

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics and Computing,

Indiana University

June 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Filippo Menczer, Ph.D.

Alessandro Flammini, Ph.D.

Yong-Yeol Ahn, Ph.D.

Christine L. Ogan, Ph.D.

Weihua An, Ph.D.

April 25, 2017

“ *Tiger got to hunt,*

Bird got to fly;

Man got to sit and wonder, “Why, why, why?"

Tiger got to sleep,

Bird got to land;

Man got to tell himself he understand.

”

Cat’s Cradle , *Kurt Vonnegut*

ACKNOWLEDGMENTS

I would like to start with a statement: I have never been in a good relationship with words. Using language appropriately to reflect how I feel and think is always a challenge for me even in my native language. Please bear with me and read with your own inner voice and imagine how I might feel while writing these pages.

I owe the greatest gratitude to my advisor, Filippo Menczer. I am very lucky to know him and to feel his endless support during all the stages of my academic career. None of these would be possible without his support and great mentorship. Here I would like to share my perspective as one of his graduate students because everyone I met already knows how great he is as an advisor and as a colleague. He is always patient with me even at 2 AM when fixing the smallest grammar mistakes I made like a father teaching a toddler how to walk. He is the most open-minded person, who believed in me and supported my *dreams*. I remember days when he boosted our motivation by inviting us for a quick foosball game. It was an honor to work with him. He set the bar really high and I feel responsible for being as good an advisor as he. I should also thank Colleen Menczer for her hospitality in the NaN Labor day parties. Without her understanding and support for Fil, some of our papers that required working late at night just before the deadline wouldn't have been submitted in their best shape.

All Ph.D. students wish to have a good advisor and I am exceptionally lucky to have two. Alessandro Flammini is a great mentor. His critical thinking and suggestions always improved our work and papers. Our motto is "Today is the day!" and by repeating this

every day we collected precious memories indicated as $|\aleph|$. His encouragement always keeps my excitement alive.

Emilio Ferrara was one of the first people with whom I interacted in our research group. I remember the day when he left Granovetter's paper on my desk like yesterday. We have been collaborating on several projects, many of these are part of this thesis. He is a great role model and his ambition for doing research provided me with an early momentum. Since our first paper together I have been following his never-ending enthusiasm.

I would like to give my sincere thanks to my research committee Christine Ogan, Yong-Yeol Ahn, and Weihua An for agreeing to serve on my committee, their constructive feedback, and support. I acquired a rich set of skills and experiences from their classroom. I would also like to thank Christine Ogan for our collaborative project on Gezi protests. In my last semester, I had a chance to develop teaching skills as a teaching assistance in Filippo Radicchi's class. I would like to thank him for creating this opportunity for me.

Prior to joining IU, I had great teachers and advisors during my education in all levels. I would like to thank one more time Nazim Yilmaz, Nurten Alpaslan, Mustak Erhan Yalcin, Ayze Erzan, Deniz Yuret, and Alkan Kabakcioglu. They had a tremendous impact on my professional development. I would also like to thank Hasan Murat Akinci and my dear friends from OTOKON student branch for providing me with the mindset to learn how to rapidly prototype and evaluate several ideas.

During my two summer internships at Microsoft Research, I had a great experience in doing research outside of my group. I had a chance to observe the great environment of an industry research lab. I especially would like to thank Emre Kiciman for his mentorship and professional advice. During my internship, I had great collaborators and I want to thank Abhimanyu Das, Sreenivas Gollapudi, and Alexandra Olteanu.

I always appreciated our friendly environment in the NaN group. Over the years we had

several alumni and graduate students. We had wonderful tea hours, foosball tournaments, lunch breaks together. I would like to thank Mohsen Jafariasbagh, Lilian Weng, Jasleen Kaur, Dimitar Nikolov, Azadeh Nematzadeh, Clayton Davis, Prashant Shiralkar, Qing Ke, and other members of the NaN. Our endeavor on bot detection wouldn't be as successful without Clayton's diverse skillset and Prashant's collaboration in the DARPA challenge.

I would also like to thank Tara Holbrook, Linda Hostetter, and Christi Pike for providing administrative help and patiently answering all my question over the years. The tech support team at SoIC, especially Rob Henderson and Dave Cooley, have provided tremendous help. I would also like to thank Kyle Thompson for assisting me on my OPT application.

During my busy working days in Bloomington, I was fortunate to have great friends to go out and have some drinks. I want to thank Hasan Kurban who is a great friend and was my witness at my wedding. I am happy that he introduced himself in the statistics class that we took together. I also deserve the honor to make him a real coffee drinker by forcing him to stop eating sugary Starbucks drinks. I would also like to thank my dear friend Murat Ozturk. In my last year in Bloomington, we shared the same house but more importantly, we engaged in very deep and intellectually stimulating conversations. I hope Murat will remember our time at IU in the future and continue our friendship since I choose to be a "poor, miserable" academic rather than going to a well-paying industry job. During my Ph.D., I have been commuting between Bloomington and South Bend. I want to thank our friends at the University of Notre Dame: Muslum and Yeter Aydogmus, Aylin Acun, and Isik Can.

I am thankful to my beloved family. They have been very supportive and I am very lucky to have them. I also want to thank my talented and wonderful sister Beril Varol, who patiently read some of my early drafts. I feel sorry that I cannot go back home often to see how you grow up to be a wonderful woman and share experiences and struggles of my family.

I would like to use this opportunity to send a Turkish message to my parents: "Bunlar daha antrenman diye diye annemin diledigi kursulere iyice yaklastim. Sizi seviyorum." I also would like to thank my parents-in-law, Salih and Hatice Mustafaoglu, for their prayers and support. Thanks to Nuray and Yavuz Cicek for our fun Skype meetings in the weekends and their company during our time in Turkey.

I do not even know how to thank my incredible wife, Nur Mustafaoglu, because I cannot imagine a version of myself that exists in a parallel universe without her. She has been everything to me. She is my motivation to work hard on the weekdays to create more time with her over the weekend. I love these lyrics by the Proclaimers "I would walk 500 miles; And I would walk 500 more; Just to be the man who walked a 1000 miles; To fall down at your door." We are more than two bodies across Indiana but one heart beating together. I love you so much ♡.

The most difficult part of completing my Ph.D. is the realization of how much I will miss my days at IU. Happiness is the most important aspect of life because I believe anything is possible when people find happiness in their work and life.

> *"I'm just one hundred and one, five months and a day."*
>
> *"I can't believe that!" said Alice.*
>
> *"Can't you?" the Queen said in a pitying tone. "Try again: draw a long breath, and shut your eyes."*
>
> *Alice laughed. "There's no use trying," she said: "one can't believe impossible things."*
>
> *"I daresay you haven't had much practice," said the Queen. "When I was your age, I always did it for half-an-hour a day. Why, sometimes I've believed as many as six impossible things before breakfast."*
>
> Through the Looking Glass, *Lewis Carroll*

* This thesis was supported in part by DARPA grant W911NF-12-1-0037, NSF grant CCF-1101743, ONR grant N15A-020-0053, and the J.S. McDonnell Foundation. I would also like to thank the School of Informatics and Computing at Indiana University Bloomington and supporters of student travel funding that I received in the past 5 years.

Onur Varol

ANALYZING SOCIAL BIG DATA TO STUDY ONLINE DISCOURSE AND ITS MANIPULATION

The widespread use of social media helps people connect and share their opinions and experiences with millions of others, while simultaneously bringing new threats. This dissertation aims to provide insights into analysis of online conversations and mechanisms that might be used for their manipulation. The first part delves into the effect of geography on information dissemination and user roles in online discourse. I study trending topics on Twitter to highlight mechanisms governing the diffusion of local and national trends. My analysis points to three locally geographic regions and one cluster that contains trend-setting cities coinciding with major travel hubs. When factors limiting information spread are considered, censorship mechanisms mandated by governments are found to be ineffective and even show a correlation with increasing popularity. I also present an analysis of spatiotemporal characteristics and distinct user roles in the Gezi movement. Next, I discuss different forms of social media manipulation. Malicious entities can employ promotion campaigns and social bots. We build machine learning frameworks that exploit features extracted from network, content, and users to train accurate supervised learning models. Our system for early detection of promoted social media trends harnesses multidimensional time series signals to reveal subtle differences between promoted and organic trends. In my research on social bots, I carried out the largest study of the human-bot ecosystem to date. Our estimates suggest that between 9 and 15% of active Twitter accounts are bots. I present distinct behavioral groups and interaction strategies among human and bot accounts. This body of work contributes to a more comprehensive understanding of online user behavior and to the development of systems to detect online abuse.

Filippo Menczer, Ph.D., Chairperson

Alessandro Flammini, Ph.D., Member

Yong-Yeol Ahn, Ph.D., Member

Christine L. Ogan, Ph.D., Member

Weihua An, Ph.D., Member

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

> *"Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."*
>
> Lewis Carroll, *Alice in Wonderland*

1.1 Motivation

Communication is a central part of society and crucial for human evolution [171]. All forms of living develop or inherit ways to interact with each other [304]. Shannon's ground-breaking work formally defines components of efficient communication systems and introduces concepts about information, noise, and bandwidth [253]. Throughout human history, we can see all forms of communication: verbal, written, and artistic expressions. Even the simplest form of communication, drawing, serves as records to communicate with future generations. The formation of signals and invention of languages are inevitable for evolving groups and systems to transfer information [261]. Over the centuries, technology helped us to develop more efficient models of communication. The invention of the telegraph and the telephone overcame the difficulty of transmitting information to distant places. These peer-to-peer communication systems mirror our natural interactions.

Each communication system consists of three main components: sender, receiver, and media for dissemination. In most cases transmission between the sender and the receiver is not perfect and this can be attributed to the noise interfering in the media or how information is encoded and decoded by the sender and the receiver respectively. It has also been observed that the sender adjusts its language and style to align with its audience [92]. Examples of language and style matching can be seen in language mimicry observed in the context of power differentials between discussants [89] and prediction of message popularity [269]. In social psychology, there has been a large body of work on persuasion and social influence [63,71,306] that talks about various cognitive theories and psychological processes behind how people convince and persuade each other. Guadagno and Cialdini discuss persuasion and compliance in the context of Internet-mediated communications, especially textual messages [144].

As the information within our reach grows exponentially, attention becomes the limiting factor in the consumption of the information. Human communication is limited due to evolutionary pressure to focus attention and use our resources efficiently [108]. Herbert Simon introduced the term *attention economy* to explain human attention as a scarce commodity and economic theory behind the various information processing strategies [260].

To overcome attention and noise limitations, we invent different modes of communication. When popularity and influence of the content are taken into account, information producers should adapt different strategies to convey their messages or use a medium that supports broader dissemination. To save time when sharing the same content, we *broadcast* to larger audiences. Broadcasting information in large-scale introduces new one-to-many channels for information dissemination. Radio, television, and newspapers are examples of one-to-many communication.

The unprecedented increase in social media use may be the result of our limited attention

and desire to reach information fast. Using the Internet, we can access vast amounts of information anytime we want. We can also prioritize, filter, and endorse relevant context. Researchers emphasize the importance of the Internet to study mass communication and how theories about communication can be applied to this new medium [211]. The Internet provides a reliable infrastructure to access and disseminate information. In this dissertation, I draw some parallels between existing theories and their correspondence in social media analysis.

The concept of diffusion is not new. We can think of diffusion as information transferred between individuals. Everett Rogers studied diffusion of innovations [244]. His groundbreaking work laid out the properties of each elements necessary for a successful diffusion system: innovation, adopter, communication channel, time, and social system.

Every communication system has a certain level of noise and disruption that impacts the efficiency of the overall system. Temporal durability of message and limited attention of the receivers may be some of the significant challenges for earlier communication systems. Recently, we have been facing more serious problems: deception, censorship, and abuse. Volume and velocity of the online data facilitate manipulation and targeting strategies toward certain groups with a higher rate of success. Researchers study these problems and develop systems to prevent unwanted consequences. Efforts to educate Internet users are also a great endeavor to prevent the dissemination of unreliable and misleading news.

Politics in broad terms can be defined as the process of making decisions that apply to all members of the groups, or alternatively, politics can be defined as the person who tries to influence the way a country is governed. To obtain such power and influence, politicians work towards obtaining trust and persuading oppositions to change their attitudes. To reach their goals, they use available technologies efficiently.

In the political system, we have been observing the impact of different communication

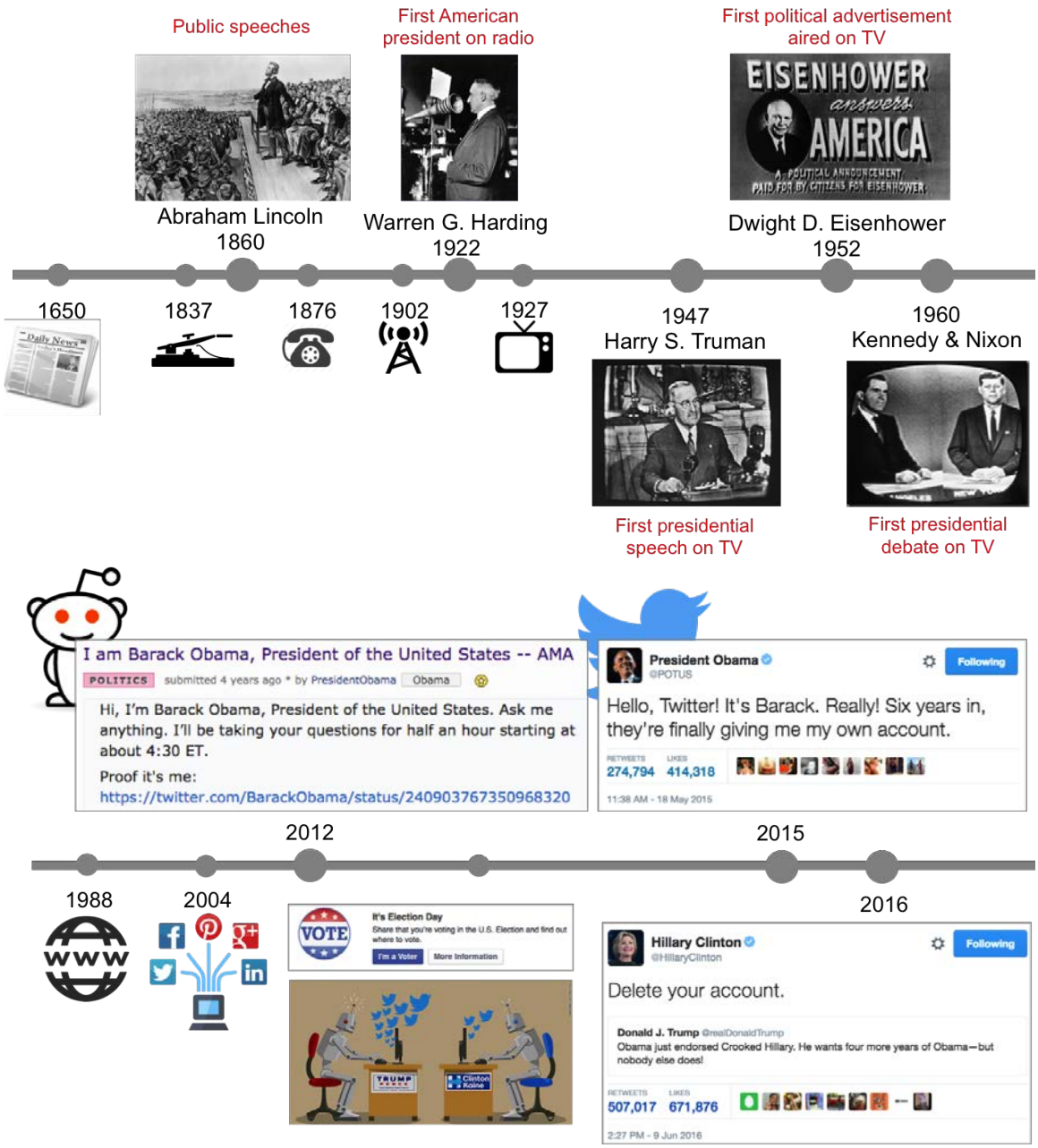

Figure 1.1: Timeline of US politics and its relation with the technological developments. Some of the key events are selected. The top panel presents influence of traditional communication media such as newspapers, radio and television. The bottom panel starts with the invention of the Web and presents some key events of US politics on the Internet.

media and how politicians adapt their strategies to influence and persuade voters and citizens [59]. We depicted a timeline representation of technological development and how politicians adopt these trends in Fig. 1.1.

In the early days, newspapers and telegraph were important to diffuse news [41, 111]. These technologies accelerated the information diffusion rate from days to hours. Organizing public speeches in parks and squares became more convenient because organizing and informing a broader audience became possible in the early 19th century. Important policy decisions and public affairs could also be shared more conveniently.

The invention of the telephone and the radio, in the 1870s and 1920s respectively, created opportunities for politicians to reach out to larger groups. Television changed political campaigns significantly [24, 259, 300]. Only ten years after the first news aired on the BBC, President Truman gave his presidential speech live on TV in 1947. This trend was followed by the first TV advertisement by Eisenhower in 1952 and the first presidential debate between Kennedy and Nixon in 1960. One estimate of President Truman's campaign indicates that he could travel more than 31k miles and meet 500k voters in person. Almost four years later, Eisenhower reached millions through television advertisements [102]. It is also important to note that those political contacts during campaigns also changed to become carefully engineered and studied.

The information age has transformed our experience in various ways. The Internet turns out to be a valuable resource to study and answer valuable questions about communication in general [211]. Politicians have become active users of the social media. They can engage with their constituents and campaign on social networks. According to an analysis by the Pew research center, 65% of US adults are actively using social media [233].

Observation and deliberate consideration of problems on social networks point to the challenging questions: What are the implications of censorship on social media? How do users behave during social upheavals? Can we detect online campaigns? How can we identify and characterize social bots? This dissertation aims at providing a systematic analysis of online discourse in terms of trend diffusion, censorship, and user behavior during a social

upheaval. Manipulation of online discourse is also studied for detecting online campaigns and social bots. The work presented in this dissertation is expected to have implications for many fields, including social media analysis, online marketing, and prevention of abuse on social media.

1.2 Research Questions and Overview

Studies of online discourse and its manipulation have great societal impact. We are using social networks and online platforms in nearly every part of our lives. We reach out for information, interact with friends and during all these processes we leave digital fingerprints about our activities and behaviors. In this work, we are interested in how underlying systems foster information diffusion for daily communication and are affected by external influences such as censorship and social protests. The dynamic landscape of online networks is vulnerable to attacks by malicious entities. Groups continuously try to influence public opinion, to pollute public discourse, and to promote their ideas. Orchestrated campaigns and social bots are ways to gain power on social networks.

In Part I, I present a study of information diffusion where geographic constraints are introduced. We analyzed information diffusion in the context of popular memes such as trends. In parallel, we analyzed social media censorship when popular or important content is prevented from reaching a broader audience as a result of governmental requests. In addition, we studied a social protest from Turkey. We characterized the role of users and how exogenous events and collective behavior affect events as they unfold.

In Part II, I discuss the important problem of identifying social media campaigns. Social media provide channels to propagate messages to people of interest. Some of this content might be promoted by advertisements and gain artificial popularity. In this work, we analyze Twitter trends and promoted hashtags to evaluate our system to detect campaigns.

Part III focuses on social bots with several goals: (i) building a machine learning framework that identifies bots with high accuracy; (ii) estimation of social bot presence on social media; (iii) characterization of human-bot ecosystem and behaviors of social bots. The next section introduces the general research questions examined in each part of this dissertation.

1.2.1 Part I: Analysis of Online Discourse

Starting from the geography of information diffusion in the context of trends and censorship, we can ask the following research questions:

- **What is the relation between geography and trends?**
- **Is censorship in a particular country sufficient to prevent diffusion of sensitive content?**

The work on trend diffusion is motivated by the observation of same or similar hashtags emerging from different geographic regions before reaching country level popularity. We identified three distinct geographical clusters in the US information flow (east coast, midwest, and southwest) as well as global patterns in the flow corresponding to main air traffic hubs. We uncovered two distinct dynamics of diffusion: localized diffusion of popular content and global spread through major hubs. We showed that travel hubs act as trendsetters, generating topics that eventually trend at the country level, then driving the conversation across the country [116]. Analysis of censorship shows that withholding content from a particular country is not sufficient to eliminate diffusion of sensitive topics. Accounts following censored discourse and censored users spread censored content by finding alternative ways to breach those geographic limitations to reach and promote content [281].

We study information diffusion and user roles during social upheaval in Turkey. We explore the following question to understand the dynamics of social protests and user roles:

- **Is online user behavior affected by external factors and do such factors cause the emergence of collective behavior?**

Our work on the Gezi protests analyzes spatio-temporal characteristics of how events unfold. We identified user roles based on their activity and involvement in information creation. Our analysis reveals that the conversation becomes more democratic as events unfold, with a redistribution of influence over time in the user population. We conclude by observing how the online and offline worlds are tightly intertwined, showing that exogenous events, such as political speeches or police actions, affect social media conversations and trigger changes in individual behavior [224, 284].

1.2.2 Part II. Detection of Campaigns

Social discourse can be controlled and manipulated through orchestrated campaigns. In this part of my dissertation, I analyzed promoted content on Twitter as a proxy for social media campaigns. Our work on campaign classification and detection addresses the following questions:

- **How well can we distinguish promoted trends from organics ones?**
- **Can we detect campaigns in their early stages on Twitter?**

In this work, we designed a machine learning framework to tackle this problem. Our supervised learning framework exploits hundreds of time-varying features to capture changing network and diffusion patterns, content and sentiment information, timing signals, and user meta-data [117, 283].

1.2.3 Part III. Analysis of Social Bots

Conversations on social media can also be manipulated by users controlled by automated scripts called bots. This part of the dissertation analyzes social bots and their behaviors in detail. We make the following contributions and answer several research questions:

- **Can we build a highly-accurate framework to detect and study social bots?**

- **What are some heuristics that we can use to detect social bots?**
- **What fraction of Twitter accounts are social bots?**
- **Can we quantify strategies adopted by social bots?**

In this work, we start building a social bot detection system called *BotOrNot*[1] and an API for other researchers to use our system [94]. Leveraging the lessons learned from BotOrNot we participated in the DARPA social bot detection challenge and we finished this competition as the second fastest and the third most accurate team [266]. Using our framework, we analyze a large-scale collection of active Twitter users to estimate the fraction of active bot population on Twitter [282].

[1]Our system will soon be renamed BotOMeter.

CHAPTER 2

Related Work

> *Nearly everything is really interesting if you go into it deeply enough.*
>
> Richard P. Feynman,

The unprecedented increase in social media use brings many opportunities and threats at the same time. Social media help people to connect and share their opinions and experiences with millions of others. We can consider social media as a microscope for the online world which magnifies individual and group behaviors. Using social media as a tool researchers can study online protests, political debates, and changes in user behaviors. The adoption of online systems has been changing the communication landscape; diffusion of online information has exceeded the limits of earlier methods of communications such as newspapers, radio, and television. These media all have an important role in information diffusion. Nowadays the Internet provides instantaneous reach to information, but it also enables the creation of misinformation. Malicious intentions can be observed in the form of orchestrated campaigns and promotion of content with the help of social bots. Detection of misinformation campaigns and social bots is crucial for our modern society. This chapter summarizes and reviews existing literature on social media studies.

Figure 2.1: Some of the notable examples of propaganda posters: “Uncle Sam" [123], “Daddy, what did YOU do in the Great War?" [81], “We Can Do It!" [275], and a propaganda poster from the USA against Nazis and Japanese during the WWII [124].

2.1 Propaganda and Campaigns on Traditional Media

Traditional communication channels like newspapers, radio, and TV changed how political campaigns have been organized and how campaign money has been spent to use those platforms most efficiently. In the introduction, we provide some examples from US politics, but these observations are applicable to most countries. Here we will delve into campaign strategies adopted on traditional media channels.

Advertisement has a significant role in reaching voters and the goal of a successful campaign is to choose the right approach to win the election. The most successful campaigns have the most memorable themes and visuals that help sway public opinion and win elections.

Persuasion is the main tool in traditional campaigns. All forms of the campaign (posters, TV ads., etc.) are the products of carefully engineered themes and messages. How public opinion is created and shaped in advertisement campaigns is explained by Edward Bernays in his seminal work “Crystallizing public opinion” with various examples [32].

Earlier engineered persuasion campaigns used printed media such as posters and newspaper advertisements to reach targeted audiences. Common themes in these posters are

depicting an enemy as evil or portraying yourself to look righteous [194]. Some of the most memorable posters target different personal traits and moral foundations as well (see Fig. 2.1). For instance, the "I Want You" poster presents Uncle Sam as a way to manifest patriotic emotion, which is used to recruit soldiers for both first and second world wars. A similar example of recruitment propaganda is released by the British government during WWI, which shows a daughter posing a question to her father, "Daddy, what did YOU do in the Great War?". This poster is trying to manipulate an able man with guilt associated with not volunteering for wartime service. "We Can Do It!" is another wartime propaganda used to boost worker morale during WWII that later became popular to promote feminism and other political issues [152, 256]. An example of a poster that demonizes the enemy is presented in Fig. 2.1.

Perhaps not surprisingly, we observe an increase in comic book sales during international conflicts [212]. Comic books are predominantly used as propaganda tools by using visual cues to present cultural ideas embodied in flesh-and-blood characters. Ideas about nationalism, societal stability, and feminism were best presented by Superman, Batman, and Wonderwoman respectively [64].

Themes and motives used in television advertisements show common parallels with the propaganda posters used during the Second World War. An analysis of over 800 TV advertising spots between 1960 and 1988 shows that *negativity* in advertisements mostly appeals to voters' fears [163]. We observe shared components such as triggering fear and emotions, nationalism, and demonizing the enemy. Tony Schwartz, a media consultant, created two of the most memorable election advertisements in US politics. "Daisy" spot were aired only once in 1964, but later replayed several times in other news outlets because of its emotional impact. In this short clip, the association between a countdown for the atomic bomb and a young girl counting daisy petals triggers emotional response and fear.

Using the power of television, politicians reach out to larger crowds and drive their attention as they choose [30,102,106,149]. Advertisements play the important role of putting the "typical citizen" on the spot and setting norms and important questions. Politicians use advertisement to make an effective campaign by either supporting their own campaigns or attacking the policies of their opponents [24, 259]. One of the first examples of this effort was known as the "Eisenhower answers America" campaign, where the President answered question recorded in a studio that contained important messages for his campaign.

Similarly, advertisements also use celebrities who have an influence on people. Creating associations between admired celebrities with certain ideologies is another strategy used in political campaigns. McAllister discussed personalization of politicians and how political priming works through television [196].

Persuasion is a broad term that covers different types of influence. We talked about how advertising is used to influence political beliefs. However, influence through advertisement is not the only type that affects and changes belief systems [72]. Most engineered persuasion campaigns contain a certain level of misinformation [258]. Fake news and conspiracy theories are examples of such campaigns.

Since ancient Greek times, rhetoric and elocution have been recognized as the highest standard for a successful politician. Aristotle's rhetoric describes three main mechanisms for persuasion: ethos, pathos, and logos [13]. *Ethos* is an appeal to an authority or credibility of the presenter. If a presenter has credibility and shares certain moral values, these moral values can be used to support message. Examples of such campaign were common during cigarette advertisements that used actors dressed as doctors mislead audiences. *Pathos* is an important component, which appeals to the emotions of audience. Pathos might use not only positive emotions like hope and gratitude but also negative emotions like fear and threats. Lastly, *logos* is the logical appeal or the simulation of it. It is commonly used with

facts and figures to support claims by the presenter. It is commonly used together with ethos.

Most persuasion campaigns use strategies that present content along with conflicted facts and distorted claims by authorities [78, 312]. Conspiracy theories are one of the most extreme but persistent examples of misinformation. They appeal to the psychological urge to explain that mysterious things happen for a reason [132, 268]. Successful conspiracy theories emerge from a group of supporters, who believe in the sinister aims of higher entities such as governments, religious groups or even extraterrestrial life forms [141].

Censorship is a practice to repress dissemination of the truth. Historically, we observed practices like collecting printed media, preventing the release of movies or manipulating pictures or news to hide facts. Censorship of various newspapers was protested by printing censored content in blank [80]. Examples of such counter-censorship tactics can be seen in French, Australian, and Palestinian media. Nazis and Stalin collected books and other printed media and burned them during political repressions [134]. Such practices inspired dystopian novels like *Fahrenheit 451* [49].

2.2 Social Media: Microscope for World

2.2.1 Memes and Trends

The *meme* concept was first proposed by Richard Dawkins in his influential book "The Selfish Gene" [95]. Dawkins defines the meme as "a unit of cultural transmission, or a unit of imitation". Nowadays we have adopted this concept to represent hashtags, keywords, and URLs on the internet.

Tracking and grouping similar concepts is easier when they are presented as quantifiable units. Memes serve this purpose. Most of the studies that analyze the content generated online isolate memes as starting points for their initial datasets.

There has been a large body of work in the area of information diffusion through networks. Several early models for information diffusion were inspired from classical disease propagation models in epidemiology, such as SIR and SIS [16]. There has also been extensive work on modeling the adoption or spread of an idea, content or product in a social network. Well known classes of models in this domain include Threshold [142] and Cascade models [135], that specify how a node adopts a particular idea or product based on the adoption pattern prevalent in its neighborhood. The concept of diffusion was initially introduced by social scientists and theory was developed to study how innovations and novelties spread [244]. Studies also define different categories for adopters such as innovators, early adopters, majority, and laggards based on their rank in involvement. Other related diffusion models for product marketing included the Bass [22] model that is based on an S-shaped adoption curve [122].

In recent work, Goel *et.al* proposes a formal measure, *structural virality*, of the degree to which a cascade reaches its audience through broadcast-like mechanisms vs. viral mechanisms [131]. The authors conduct a large scale empirical study of a billion diffusion events for news, videos, images and petitions on Twitter and observe a wide range of diverse cascading structures with varying structural virality, and show a low correlation between popularity and structural virality. The authors then show how a simple SIR model can capture several of the empirically-observed properties of the cascades. However, they note that their model could not explain the large variance in structural virality that they observed empirically.

Trends represent interesting collective communication phenomena: they are user-generated, continually changing and mostly ungoverned (although orchestrated hijacking attempts have been observed [52, 240, 241]). Different information diffusion mechanisms may determine the trending dynamics of hashtags and other memes on social media. Exogenous and endogenous dynamics produce memes with distinctive characteristics [116, 119, 181, 216, 263]: external

events occurring in the real world (e.g., a natural disaster or a terrorist attack) can generate chatter on the platform and therefore trigger the trending of a new, unforeseen hashtag; other topics (e.g., politics or entertainment) are continuously discussed and sometimes a particular conversation can garner lots of attention and generate trending memes. So far, trends have been studied as a proxy to detect exogenous real-world events discussed in social media [5, 23, 87, 246], emerging topics, or news of interest for the online community [60, 183].

Recent work analyzes emerging topics, memes, and conversations triggered by real world events [5, 23, 60]. Studies of information dissemination reveal mechanisms governing content production and consumption [73] as well as prediction of future content popularity. Cheng *et al.* study the prediction of photo-sharing cascade size [65] and recurrence [66] on Facebook.

2.2.2 Geography of Information Diffusion

It has been suggested that social media may overcome the spatio-temporal limitations of traditional communication: technologically-mediated systems make it possible to ignore physical and geographic distances [75, 217]. This, however, does not imply that communication patterns on social media are not affected by physical distances and geographic borders [209, 227].

Trends are also strongly localized in space and time: the temporal and geographic dimensions play a crucial role to determine the success of a trend in terms of spreading and longevity. Unveiling the spatio-temporal dynamics that drive trending conversations on social media is instrumental for many purposes: from designing successful advertising campaigns, to understanding virality and popularity that characterize some topics. Recent studies took advantage of platforms such as Yelp and Foursquare, which provide customized services to their users based on their physical location (*e.g.*, recommendations of events or places), to study geographic user activity patterns [221, 247–249]. Others have used plat-

forms such as Twitter and Facebook, that enrich user profiles with geographic information and accompany user generated content with location-based data, to map user demographics [174, 206].

Onnela *et al.* [227] noted that, although the probability of observing a tie between two individuals in a social network (in that case, a mobile phone call network) decreases as a power law with physical distance, the geographic spread of social groups quickly increases with the size of the group; even groups of modest dimensions ($\approx$ 30 members) are able to span hundreds of kilometers, suggesting that, in technologically-mediated social systems, there exist distinctive social dynamics that govern the communication among individuals. Geographic locations and physical distances have been found to be correlated to friendship behaviors in online social networks [187], to determine patterns in human mobility networks [51, 137], and to affect collaboration schemes in science networks [228].

Geographic factors have also been recently found to be crucial in the adoption of languages and dialects [209], and in the expression of sentiment [207, 236, 237] in online social media. Mocanu *et al.* [209] showed how social media data can be used to characterize language geography at different levels of granularity, to highlight patterns such as linguistic homogeneity and linguistic mixture in multilingual regions.

2.2.3 Proxy to Analyze Human Behaviors

Studies by Mitchell *et al.* suggests that the adoption of online social media content can be instrumental to describe emotional, demographic, and geographic characteristics of users of these socio-technical systems; in particular, they investigated Twitter users active in the US in terms of happiness and individual satisfaction [125, 207]. A study of happiness on Twitter led to a hedonometer project, in which the authors study temporal changes of global happiness and the relation between local low and high points with real-world events [107].

People use social media to reflect their emotions and events affecting their lives through social media. The mismatch between the social representation and real state of the user can pose challenges for research that leverages social media data because many individual worries about their online representations and conform online norms [109, 151]. However, our behaviors on social networks still carry a lot of information about personality, cultural, political and sexual preferences [133, 238, 239].

The use of social media also shows strong correlation with public health measures [98, 230]. Researchers have been studying several health related topics using social media data [85, 96, 97, 99, 225, 226] and search logs [229, 231, 301].

Similarly, services for online shopping have rich information about our preferences and tastes. We use health monitoring devices to track our work-out routines and sleep quality [143, 200]. Location based services like *Foursquare*, *AirBnB* and *Yelp* track our eating habits and navigation history [1, 167, 201, 257].

2.2.4 Detection of Emerging Topics

Another recent research line related to our work is that of the detection of emerging trends, topics, memes, and events in online social networks and social media [5, 23, 60, 87, 114, 183, 195, 246].

Social media data can be used to make educated guesses on the outcome of real-word events, such as elections or competitions [104]. Ciulla *et al.* [75] combined trends and geographic information of Twitter data to demonstrate that online social media can be exploited to predict social events in the real-world. They collected trending hashtags and phrases related to contestants of the popular TV show *American Idol,* mapping the fan base of each candidate to different geographic regions inside and outside the US, to identify spatial patterns in attention allocation and preferences expressed on the online platform.

These signals were then combined and used to predict voting behaviors of fans achieving good accuracy.

2.3 Social Media: Online Discourse Platform

Technologically mediated communication systems, like social media platforms and online social networks, support information sharing and foster the connectivity of hundreds of millions of users across the world every day [48, 288]. The adoption of these platforms has been associated with profound changes in 21st-century society: they affect how we produce and consume information [11,44,218], shifting the paradigm from a broadcasting model (one-to-many, like radio and TV) to a peer-to-peer (many-to-many) distribution system. They have also altered the ways we seek information to understand societal events surrounding us [203, 204], and how we interact with our peers [61, 62].

People participate in social media for many different reasons. Some join social media with the intention to socialize with friends or to meet new people. Others participate to promote a cause, or to gain fame as an authority or expert in their topics of interest. Much prior research has documented the many reasons why people choose to participate on social networks, such as communicating with real-world friends or making new contacts [162, 177], connecting with colleagues and building professional relationships [105, 168], and connecting with users that act as information providers [158].

2.3.1 Social Media Use During Protest

Ease of access to online services creates opportunities to freely discuss and share opinions and to debate different points of view. Political discussions are the most influential for individuals and consequential for society. Recently, researchers have been studying political uprisings and social protest around the world using data collected from various online platforms.

Examples of social protests and movements that have used social mobilization include the revolution in Egypt [68], the anti-capitalist Occupy Wall Street movement [56,67,83,84,101], and social upheavals in Spain [45, 139] and Turkey [53, 155, 224, 284]. The benefits resulting from the adoption of social media include lowered barriers to participation, increased ease with which small-scale acts can be aggregated, the rapid propagation of logistic information and narrative frames, and a heightened sense of community and collective identity [28, 29, 214, 279, 307]. These events provide evidence of the impact of social media and their importance for free speech. Protecting these resources from disruption is important for the continuous free flow of information in society.

Social media have played a pivotal role in the development and increasing frequency of social movements [28, 128, 214]. Using survey methodology, Tufekci and Wilson [271] found that the use of social media in the Egyptian protests allowed people to make informed decisions about participation in the movement, provided new sources of information outside of the regime's control, and increased the odds that people participated in the protests on the first day. Another survey found Facebook use for news and socializing in Chile's youth movement to be positively associated with participation in the protests [278]. Chief among social platforms used for protests is Twitter that, with more than a half billion users, provides a high-visibility window on real-world events and an active forum for discussion of political and social issues. The mostly ungoverned nature of this platform ensures a democratic, peer-to-peer discussion, aiming at both creating a framing language to set goals for the protest, and as a vehicle for mobilizing resources and social capital to sustain it [3, 83, 153, 190]. Individuals and organizations can discuss and share information on Twitter about the movement's political and social objectives [26, 27]. They can also coordinate to marshal the resources needed to carry out on-the-ground activities like encampments or marches [159, 198].

González-Bailón *et al.* [139] collected a large corpus of tweets related to the Spanish social and economic 'Indignados' protest that unfolded during May 2011. Their work provides evidence that Twitter played a role in the recruitment of new individuals to the protest movement as well as in the dissemination of information related to mass mobilization activities.

Choudhary *et al.* [68] analyzed the aggregate tweet sentiment during the 2011 Egyptian revolution, observing that fluctuations in positive and negative sentiment were closely correlated with the sentiment expressed by influential users worldwide. The authors also observed that users tweeting about the Egyptian revolution were distributed both inside and outside Egypt. Baños *et al.* [20, 21] highlighted the role of social media users in the diffusion of information related to mass political mobilizations, unveiling the presence of hidden influentials who foster large cascades. The authors also observed how the topology of the communication network during such events reflects underlying dynamics like information diffusion and group emergence.

In a study of the Occupy Wall Street uprising, Conover *et al.* focused on the geospatial characteristic of the protest [83]. They observed that highly-localized discussions mirrored individual attempts to organize and coordinate mobilization on the ground. Interstate discussion channels driving long-distance communications fostered the collective framing process that imbues social movements with a shared language, purpose and identity. A longitudinal analysis [84] revealed that users did not change their connectivity, interests and attention patterns with respect to baseline activity prior to the beginning of the protest. These findings left open the question whether Occupy had any long-lasting effect on its online community of participants.

2.3.2 Censorship

In some societies, governments have responded to the growing phenomenon of political mobilization by either terminating access to the online services or developing laws to restrict the exchange of information [313]. China, Iran, North Korea, and Turkey are examples of countries applying internet censorship widely. These countries are monitoring social media and news to control online discourse. If discussions turn to sensitive topics, concerned governments intervene and attempt to control information dissemination [7, 170].

Platforms like Facebook and Twitter have been censored by limiting internet access at the country level. Social media companies have recently created specialized legal departments to address requests from governments and provide continuous service for their users in censored countries. Periodical transparency reports are released by technology companies like Facebook,[1] Twitter,[2] Microsoft,[3] and Google.[4] These reports contain the statistics of requests received from different governments and disclosures of information released. The increasing trend in government requests for disclosure of user information and censorship requests are worrisome.

Many censorship regulations are developed to control or limit dissemination of political discussions. A recent study highlights a significant rate of content removal on Weibo [19]. When compared to the volume of politically relevant keywords, the authors estimated that 16% of political posts were deleted by authorities on Weibo. The content analysis of censorship on Weibo points to the discrepancy between the Chinese Communist Party and the oppositions [289]. The lag time between content creation and censorship indicates a distinction between relevant and dangerous topics from the viewpoint of Chinese censorship. The political impact of micro-blogging platforms is analyzed by comparing Twitter and Weibo

[1] govtrequests.facebook.com
[2] transparency.twitter.com
[3] microsoft.com/en-us/about/corporate-responsibility/reports-hub
[4] google.com/transparencyreport

use in China [267].

In another analysis of Weibo, researchers studied the mechanism of Weibo's trending topic detection system to track sensitive viral discussions [315]. The authors also showed the mechanism behind filtering by tracking sensitive users [316]. They found that the trend of a viral topic is short-lived, which points to the effectiveness of Weibo's censorship on sensitive topics. We also observed a small but significant decrease in median censorship time. On Twitter, censorship requires legal documents and process time unlike Chinese social media, which has centralized control of censorship.

Technical challenges against censorship can be supported by using technologies like VPN services or TOR project.[5] Researchers also built services to quantitatively measure the censorship problem [54] and analyzed examples of country-wide Internet outages [88, 287].

2.4 Social Media: Medium for Abuse

Through social media platforms, we are exposed to a tremendous amount of information. Still, we have been facing a significant problems of misinformation [113] and trapping inside an echochambers [2, 82]. Some governments are also taking precautions by applying censorship to the Internet, which eventually terminates users' right to access information.

2.4.1 Misinformation and Manipulation

Individuals and their opinions are increasingly influenced by information spreading on social media. Twitter, among others, conveys hundreds of million messages per day and plays a crucial role in the timely diffusion of news and information. Examples of Twitter conversation topics include coordinated social mobilization [84,139] and political debates [44]. Social media content is mostly ungoverned and therefore it can be manipulated. This often results

[5] `torproject.org`

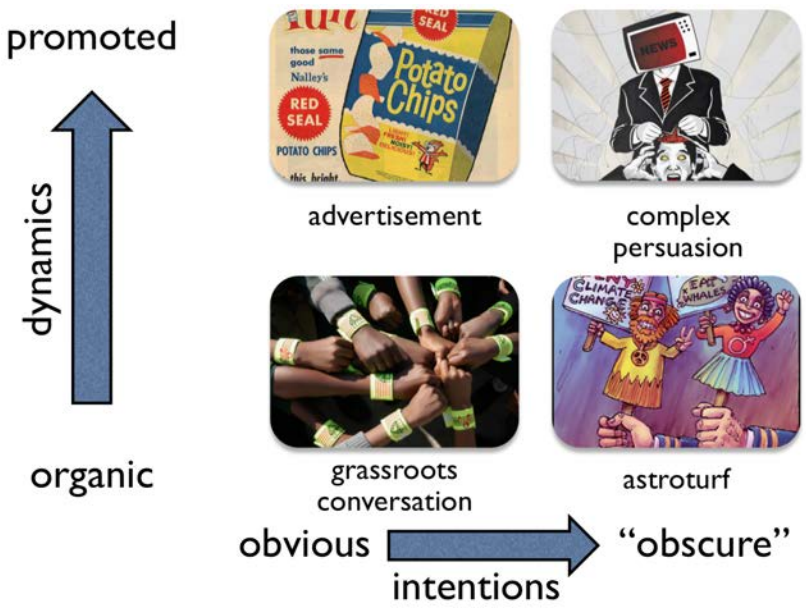

Figure 2.2: Persuasion defined according to the mode of propagation and the entities behind it.

in the diffusion of spam, misinformation, rumors, and deceptive messages [176,235]. Persuasion campaigns and other types of engineered social media conversations aim at challenging or changing reader beliefs, opinions, or ideas. The appearance of an organic movement can be created variations of the original message. Artificial means like social bots or fake accounts can be used to rebroadcast such variants [154]. When a deceptive message produced this way is believed to be genuine by real users, it can spread virally and reach a large audience [50]. Detecting engineered or artificially sustained communication in its early stage is therefore of paramount importance to avoid deception at scale.

Figure 2.2 illustrates two dimensions along which we can distinguish between different classes of conversation observable on social media: the mode of information diffusion and the entities behind the conversation.

Persuasion campaigns can be enacted by promoting content, typically by advertising. This way the content will have higher visibility and reach. This is in contrast to organic diffusion, which stems from spontaneous collective attention toward a topic. Persuasion

can also occur by employing artificial agents using fake or compromised accounts, including social bots, to give the impression that real people are paying attention to a topic or person. Based on these two dimensions, we can identify three classes of persuasion campaigns, each distinct from grassroots conversations.

Astroturf is a peculiar form of persuasion often observed in social media in the context of politics and social mobilization [241]. It aims at simulating a grassroots conversation through an orchestrated effort. The entity who attempts to generate such orchestrated campaigns generally exploits fake accounts or social bots [154, 290]. These artificial means allow the generation of a large volume of content and simulate the online activity of real users. Cases of massive astroturf campaigns have been observed during political races such as the US senate [213] and presidential elections [204].

The use of **advertising** is possibly the most common form of persuasion in social media. The intent to promote is transparent. This method aims at attracting attention toward a given entity (e.g., a brand). Advertising campaigns are an ideal use-case for our study, since promoted content is clearly labeled as such on Twitter.

Complex persuasion campaigns employ a mix of the patterns discussed above. Complex campaigns might exhibit a mixture of promoted and organic content. Conversations that start as promoted might pick up audience attention and mutate into organic topics of discussion. Alternatively, the spark to initiate a conversation might occur naturally and later involve social bots that engage in the discussion. Complex persuasion campaigns may have large societal impact if not detected early: successful campaigns can affect users by the thousands.

2.4.2 Social Bots

As opposed to social media accounts controlled by humans, bots are controlled by software, algorithmically generating content and establishing interactions. While not all social bots are harmful, there is a growing record of malicious applications of social bots. Some emulate human behavior to manufacture fake grassroots political support [35, 240], promote terrorist propaganda and recruitment [31, 118], manipulate the stock market [113], and disseminate rumors and conspiracy theories [34].

Discussion of social bot activity, the broader implications on the social network, and the detection of these accounts are becoming central research avenues [46, 113, 115, 180]. The magnitude of the problem is underscored by a social bot detection challenge recently organized by DARPA to study information dissemination mediated by automated accounts and to detect malicious activities carried out by these bots [266].

Also known as sybil accounts, social bots can pollute online discussion by lending false credibility to their messages and influence other users [6]. A recent study shows to what extent automated systems produce content and dominate discussions about electronic cigarettes on Twitter [76]. Social bots also vary greatly in terms of their behavior, intent, and vulnerabilities. A recent study proposed a categorization scheme for bot attacks on social network [208].

Most of the previous work on detecting bot accounts has operated from the perspective of the social network platform operators, *i.e.*, with full access to all data. These techniques focus on large-scale data to either cluster behavioral patterns of users [292] or classify accounts using supervised learning techniques [180,311]. For instance Beutel *et al.* decomposed event data in time, user, and activity dimensions to extract similar behaviors [37]. These techniques are useful to identify coordinated large-scale attacks directed at a common set of targets at the same time, but accounts with similar strategies might also target different

groups and operate separately from each other.

An alternative approach to study social bots and sybil attacks is to understand what makes certain groups and individuals more appealing as targets. Wald *et al.* studied the factors affecting the likelihood of a users being targeted by social bots [291]. This approach points to effective strategies that future social bots might develop.

Structural connectivity may provide important cues. However, Yang *et al.* studied large-scale sybil attacks and observed sophisticated sybils that develop strategies for building normal-looking social ties, making themselves harder to detect [311]. Some sybil attacks analyze the social graph of targeted groups to infiltrate specific organizations [110]. SybilRank is a system developed to identify attacks from their underlying topology [57]. Alvisi *et al.* survey the evolution of sybil defense protocols that leverage the structural properties of the social graph [10].

Social bot detection tools use learning models trained with data collected from human and bot accounts. Chu *et al.* built a classification system identifying accounts controlled by humans, bots, and cyborg accounts [69, 70]. Wang *et al.* analyzed sybil attacks using annotations by experts and crowd-sourcing workers to evaluate consistency and effectiveness of different detection systems [293]. Clark *et al.* labeled 1,000 accounts by hand and found natural language text features to be very effective at discriminating between human and automated accounts [77]. Lee *et al.* used a honeypot approach to collect the largest sample of bot accounts available to date [180].

2.4.3 Fake News

The term *fake news* is not new, but the prevalence of fake news in social media introduce serious problems. Fake news websites deliberately publish hoaxes, propaganda, and misinformation pretending to be legitimate news sources. Unlike satirical news, they often aim

to mislead readers in exchange for political and financial gain.

A large amount of misinformation spreads online and its prevalence and persuasiveness can affect serious decisions around vaccination [55, 166, 223], elections [8] and stock market behavior [58, 179] among other issues. A recent study suggests that misinformation is just as likely to go viral as reliable information [254]. Dissemination of fake news is promoted by copycat websites. Once an untrustworthy source releases some content online, those copycat websites duplicate the original content. Corrections on the original source are no more relevant and useful since many other media outlets are already affected. We can make an analogy between dissemination of fake news through multiple media outlets to a disease spreading in groups.

Recent research efforts focus on modeling the diffusion of misinformation [34, 36, 100, 127, 160]. Algorithmic efforts for detecting rumors and misinformation are also crucial to prevent the spread of campaigns with malicious intents [117, 205, 235, 242].

To hinder the dissemination of fake news, both journalists and readers have great responsibilities. Online websites like *FactCheck*[6], *PolitiFact*[7], and *Snopes*[8] provide fact-checking services to debunk fake news. Fact-checking provided by online services influences opinions of voters and provides a guide to politicians in judging what news might be fake before disseminating them [126, 222]. To automate fact-checking, researchers are working on designing systems that can evaluate the credibility and truthfulness of claims [74, 309].

The problem with fake news can be partially resolved by educating Internet users. *News literacy* is important and everyone should learn how to detect fake news. Recently, we have been observing a growing community of fact-checkers. *Poynter* is one of these organizations, which has released "International Fact-Checking Network fact-checkers' code of

[6] `factcheck.org`
[7] `politifact.com`
[8] `snopes.com`

principles"[9] to promote excellence in fact-checking. Another noteworthy example is *First draft*.[10] These organizations not only provide fact-checked information about popular claims but also monitor political campaigns and elections. Collaboration between different fact-checking organizations is promoted by proposing an integrated system to share fact-checking information.

2.5 Perspective for Designing Better Systems

The influence of external factors on the US presidential election in 2016 was a controversial topic. Recent research shows evidence supporting the involvement of social bots in political discourse [35]. Bessi *et al.* estimate nearly 15% of the accounts and 20% of the tweets having involvement by social bots from both sides of the political spectrum. The participation of social bots in political conversations does not necessarily need to be sophisticated. Social bots are also known as disrupting conversations by flooding content to a particular conversation channel. The pollution of conversation on social media makes it intractable for humans looking for useful information. An example of such channel disruptions was observed in Mexico recently [265], where different hastags are flooded by social bots and force people to move discussion to alternative channels.

Through social media and anonymity, targeted attacks are possible in orchestrating a large army of social bots, trolls [199] and bullies [25, 243]. Examples of extremist activities on social media have alarmingly increased and many platforms take precautions for early-detection and prevention of such activities. Recent studies also point to social media use for recruitment to terrorist organizations on social media [31, 118, 193].

We have been observing the societal impact of fake news. An increasing number of online news websites and social networks are producing implausible content. The production of

[9] `poynter.org/fact-checkers-code-of-principles/`
[10] `firstdraftnews.com`

fake news has been increasing, but the major problem is the consumption of fake news articles. It is valuable to understand the roots of the problem before proposing solutions.

Herbert Simon's work on attention economy might explain some of our mental shortcuts; we tend to believe a content based on our opinions about our friend who shared the content. Confirmation bias is considered one of the factors [220]. According to this hypothesis, people tend to believe and seek information supporting their initial opinions. However, people tend to believe nonpartisan opinions.

Traditionally people access credible information through legitimate sources. However, in the Internet age, popular users have a stronger influence on widely consumed information sources. Most of the news articles follow a journey starting from their original source to copycat websites, social media accounts, and finally to their readers. These long chains of content cause some problems, for instance corrections made on original articles rarely propagates to the latest venue. News consumers are also not aware of the original source. Researchers studied when readers pay attention to the source of content [165]. They found that users tend to believe the content considering only the source from which they obtained news unless the subject is really important to them. This problem can be solved by focusing on news literacy. When educated online users can access fact-checking tools, it is possible to stop fake-news.

There are significant efforts to preserve the social ecosystem. Researchers develop tools like BotOrNot[11] [94, 282] to detect social bots on Twitter, Hoaxy[12] [254] to study dissemination of fake news, and TweetCred[13] [145] to evaluate credibility of tweet content. The Jigsaw lab of Alphabet has also devoted significant efforts to tackle some of the global security challenges.[14] They design systems and tools to prevent censorship and online ha-

[11] `truthy.indiana.edu/botornot`
[12] `hoaxy.iuni.iu.edu`
[13] `twitdigest.iiitd.edu.in/TweetCred`
[14] `jigsaw.google.com`

rassment. Considering the impact of technology on the dissemination of misinformation, we share a great responsibility to work together. Computer scientists, social scientists, journalists and industry partners must collaborate to implement policies and systems against online threats.

CHAPTER 3

Concepts and Methods

> *Give me a place to stand and with a lever I will move the whole world.*
>
> Archimedes,

3.1 Twitter Data

This dissertation presents studies that use datasets collected from Twitter. In our lab, we have elevated access through Twitter Streaming API,[1] which approximately includes a sample of 10% of the public tweets. Our lab also built a service, called *OSoMe*, to share derived data with researchers and citizen scientists [93].

Twitter is a popular micro-blogging platform that is available to millions of people all over the world. Users can interact by creating social ties (friend/follower relations), retweeting content of others to disseminate content among their followers, and mentioning other users (using @ sign before a username, for instance, `@onurvarol`) in their posts. Twitter users can post up to 140 characters per tweet including URLs for external media content, such as pictures and videos, alongside text. Hashtags — keywords preceded by # sign — included in tweets are used as keywords to summarize a discussion topic or to convey a message in a shortened format.

[1] http://dev.twitter.com/docs/streaming-apis

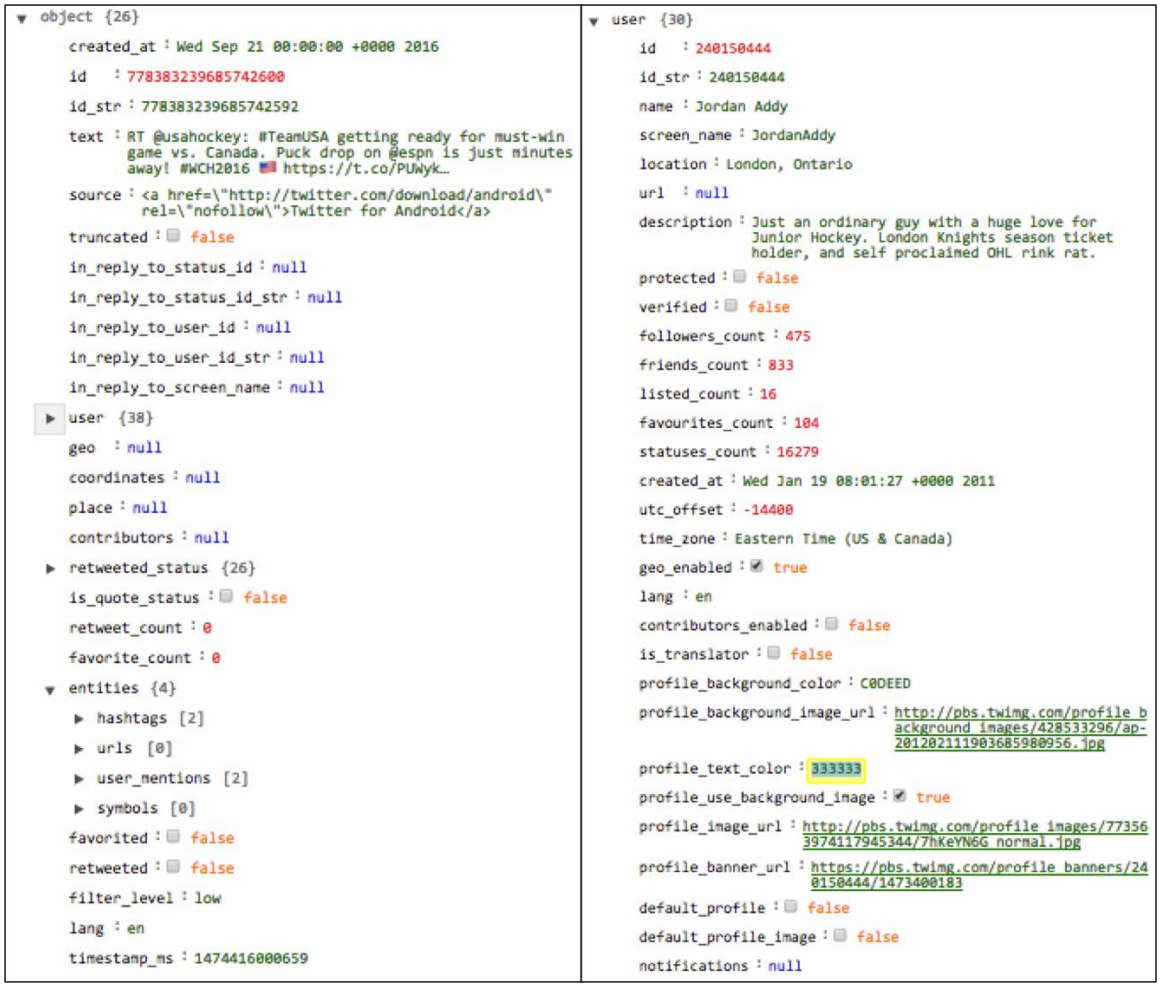

Figure 3.1: JSON hierarchy of `tweet` and `user` objects.

Twitter has a rich API to provide access information about local and global *trends*, user social network, and several other meta-data as shown in Fig. 3.1. Some of this additional information has been used in this dissertation such as trending topics for studying the spatiotemporal nature of information diffusion and *withheld* tweets for analysis of censorship on Twitter. Censored tweets contain the fields `[withheld_scope]` and `[withheld_in_countries]` to indicate how content is censored and where censorship is active.

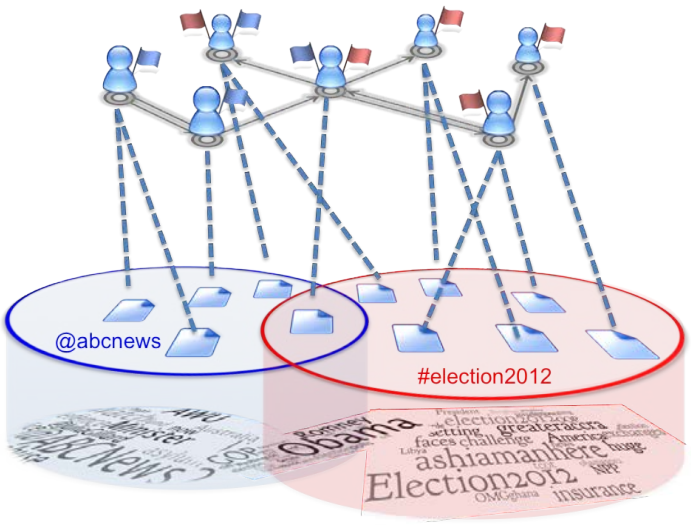

Figure 3.2: Relation between users through friendship or information flow. Posts produced during the process can be related to each other through co-occurrence or topical relevance.

3.1.1 Information Mining

Using data available on Twitter, we can extract information about the temporal evolution of user properties such as number of friends, followers, and tweets posted. We can also construct relationships between those users through retweet and mention ties. In terms of conversations, we can construct co-occurrences of hashtags and compute volume of activity. All this information, depicted in Fig. 3.2, can be extracted from a collection of tweets, which are essential to build systems described in this dissertation.

In the case of social protest or topically focused events, we can collect data through the Twitter Streaming API by keywords of interests. Users on Twitter adopt hashtags to promote communication about the events. For instance during the Gezi movement the `#direngezi` hashtag was used commonly. We collected live stream during events to learn other relevant keywords to expand our set of hashtags. Tweets that contained these hashtags were later collected from our 10% stream.

3.1.2 Feature Extraction

Using the information obtained from tweets, we can extract several features in different categories: network structure and information diffusion, language and sentiment, temporal, and user meta-data. Systems described in this dissertation use subsets of these features collected separately for each user or campaign.

3.1.2.1 Network

Twitter actively fosters interconnectivity. Users are linked by means of follower/followee relations. Content travels from person to person via retweets. Tweets themselves can be addressed to specific users via mentions. The network structure carries crucial information for the characterization of different types of communication. In fact, the usage of network features significantly helps in tasks like astroturf detection [240]. Structure and modularity of networks are also shown to be useful to maximize information dissemination [219]. We can construct three types of networks: (i) retweet, (ii) mention, and (iii) hashtag co-occurrence networks.

Retweet and mention networks have users as nodes, with a directed link between a pair of users that follows the direction of information spreading — toward the user retweeting or being mentioned. In tweet meta-data, information about the user posting the tweet is presented in the `[user]` field. If a tweet is retweeted, the original tweet is preserved in the `[retweeted_status]` field. In case of mentions and replies, tweets pointed to users can be accessed via using `[entitites][user_mentions]` and `[in_reply_to_user_id]` respectively.

The hashtag co-occurrence network has undirected links between hashtag nodes when two hashtags have occurred together in a tweet. In each tweet, the `[entities][hashtags]` field contains hashtags used in the tweet.

All networks are weighted according to the number of interactions and co-occurrences. Using these networks, several network statistics such as the number of nodes and edges, density, and average clustering coefficient, can be computed. Node-specific properties such as strength, clustering coefficient, and centrality measures can be studied through their distributions.

3.1.2.2 Language and Sentiment

Many recent papers have demonstrated the importance of content and language features in revealing the nature of social media conversations [47, 90, 184, 197, 209]. Textual information in the tweet is located in the `[text]` field. Users can also provide some free-text content about themselves such as a description and location in the `[user][description]` and `[user][location]` fields, respectively.

From the tweet texts, we extract language features by applying the *Part-of-Speech* (POS) tagging techniques using the NLTK package [39], which identifies different types of natural language components. Additional language and content features such as length of the text, number of words, URLs, mentions, and hashtags can be extracted from a tweet. User language is also available in the `[lang]` field by ISO language codes.

Sentiment analysis is a powerful tool to automatically describe the attitude or mood of an online conversation. We adopt several sentiment extraction techniques to generate various sentiment features, including *happiness score* [172], *arousal, valence and dominance scores* [295], *polarization and strength* [305], and *emoticon score* [4].

3.1.2.3 User Meta-data

User meta-data is crucial to classify communication patterns in social media [115, 206]. In a tweet the `[user]` field contains information about a user such as number of friends, followers, and posts, profile image, profile description, user language, and time-zone. Temporal changes

of the friend, follower, and post counts provide valuable information about user behaviors.

3.1.2.4 Temporal

The temporal dimension associated with the production and consumption of content may reveal important information about campaigns and their evolution [129]. Each tweet contains meta-data about creation time of the tweet and user account in the `[created_at]` and `[user][created_at]` fields respectively. The most basic time-related feature that can be considered is the number of tweets produced in a given time interval. Inter-event time distributions also carry important signals about the progression of events.

3.2 Graph Theory

The structure of a network is commonly depicted as a *graph*. A graph consists of mainly two sets of objects: nodes ($N = \{n_1, \cdots, n_N\}$) and edges ($E = \{e_{ij} | i, j \in V\}$). Each edge is represented as a pair of nodes and directionality is important if the graph is directed. Connectivity of the graph can also be represented as an adjacency matrix A, in which each element of the matrix A_{ij} represents the weight of an edge between nodes n_i and n_j. For instance a graph representation in Fig. 3.3 has 6 nodes and 9 directed edges. I use this toy-example to describe some of the important network measures below.

Degree: A number of edges connected to a node represents the degree. In directed networks, incoming and outgoing edges differ and one denoted as in- and out-degree. For instance node-3 in the network in Fig. 3.3 has in-degree 1 and out-degree 2.

Weight: The weight of an edge represents importance or strength of relation between nodes. The weight of the edge e_{ij} can be represented as w_{ij}. For instance in the toy-network the weight of the edge e_{31} is $w_{31} = 5$.

Strength: It is the sum of the weights of all edges adjacent to a node n_i. The strength

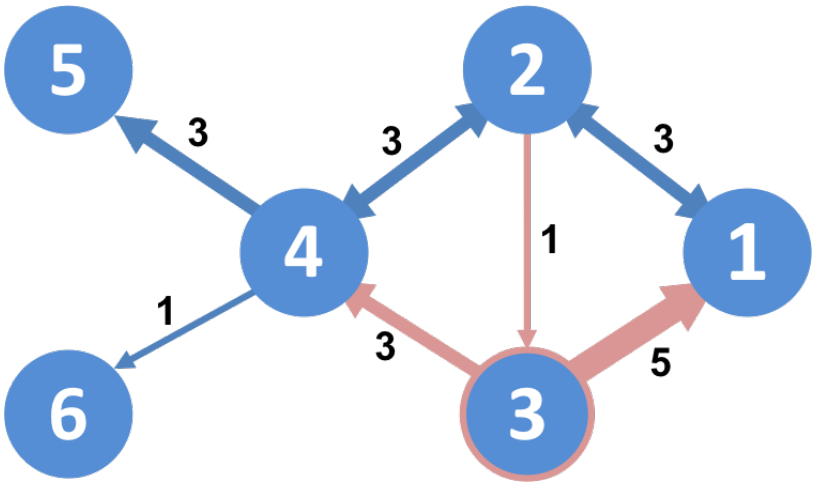

Figure 3.3: Example representation of a graph consisting of 6 nodes and 9 directed edges. Node-3 is highlighted to provide examples in definitions.

of node can be denoted as in- and out-strength if the network is directed. For instance the in-strength of node-3 is 1 and the out-strength is equal to 8.

Density: The density of a graph represents the fraction of edges that exist in the graphs compared to the maximum possible number of edges. Complete graphs have a density 1 and our example graph has density 0.3. For directed graphs the density is computed as follow:

$$d = \frac{|E|}{|V|\left(|V|-1\right)}$$

Clustering coefficient: It is a measure of the degree to which nodes in a graph tend to cluster together. There are two versions of this measure: global and local clustering coefficients. The global clustering coefficient is the average of the local clustering coefficients across all nodes in the network. Local clustering coefficient computes how close a node's neighbors are to being a clique for each node. For instance, in our toy-network node-3 has 3 neighbors. Among node-3's neighbors, 2 out of 3 possible edges are realized, which yields 0.66 clustering coefficient for node-3.

3.3 Machine Learning Methods

Every second of our lives, we observe the world and make assumptions and predictions about our environment. When these subsequent events and observations are recorded, we can also automatize these processes. We approximate the processes that explain the data by constructing models. These models might not be able to describe processes completely, but they might be useful and accountable to detect certain patterns and regularities. Machine learning explores such methods and techniques that can learn from and make predictions on data.

Techniques employed in machine learning tasks are typically classified into three broad categories: supervised, unsupervised, and semi-supervised [9]. In this dissertation, we commonly used supervised techniques to classify and detect particular patterns in the data and unsupervised techniques to explore and identify instances with high similarities.

In the following, I will describe off-the-shelf methods and evaluation techniques used in different part of this dissertation. The Python library Scikit-learn is used to apply most of these methods [232].

3.3.1 Supervised Learning

In supervised learning problems, data are presented in a form of a set of features extracted from the dataset and a corresponding label. Given a set of N training examples of the form $\{(x_1, y_1), \cdots, (x_N, y_N)\}$, x_i is a feature vector of the i-th example and y_i is its label. A learning algorithm seeks to find a function that maps the input space X onto the output space Y. In terms of binary classification, relationships between features of the training dataset are learned to map any instances of the test examples to a binary value $\{0, 1\}$ as predicted label.

Random forest is the most commonly used technique for classification and regression

tasks in this dissertation. Using a random subset of training data and a bagging mechanism (random selection of feature subset) several decision trees are constructed by algorithm [150]. Classification decision for a test instance relies on voting between generated decision trees.

3.3.2 Unsupervised Learning

Unsupervised learning aims to infer a function that describes hidden structure from unlabeled datasets. Clustering is one of the most common approaches to describe features presented in the data, by grouping data points with relates characteristics.

The *hierarchical clustering* method, which uses agglomerative clustering, is a commonly used unsupervised technique in this work. Agglomerative hierarchical clustering uses a "bottom-up" approach: each observation starts in its own cluster, and pairs of similar clusters are merged as one moves up the hierarchy. The resulting clustering of the data is usually presented in a dendrogram.

3.3.3 Evaluation Techniques

In most of the analysis with data, evaluation is a critical part of the experimentation. In supervised learning, algorithms produce predicted labels $P = \{p_1, p_2, \cdots, p_N\}$ for instances in the dataset for testing against ground-truth labels $Y = \{y_1, y_2, \cdots, y_N\}$. To overfitting noise in the training dataset, the classifier is evaluated using a scheme called *cross validation*. This common practice requires splitting data into k different folds; classifiers are trained using data in $k-1$ folds and tested on the remaining fold; this process repeats k times until each fold is used for testing. Average results of these k-fold are reported in the experiments.

The outcomes of any binary classification task can be presented in a 2x2 table called *confusion matrix*. Columns and rows of this matrix represent numbers of items in predicted and true conditions. Terms *true positive* (TP) and *true negative* (TN) represents when

		Predicted label		
		p	**n**	**total**
Correct label	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Table 3.1: Example confusion matrix representation.

predicted and true labels are matched (see Fig 3.1).

Using this confusion matrix representation, we can define measures to evaluate the performance of the classifier. Some of the commonly used measures are:

Accuracy: Fraction of correctly labeled items among all test instances:

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

Precision: Fraction of positive classifications that are correctly classified:

$$precision = \frac{TP}{TP + FP}$$

Recall: Fraction of positive instances that are correctly classified:

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

F_1 **score**: Harmonic mean of precision of recall:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

AUC: Area under the received Receiver Operating Characteristic (ROC) curve is a measure of accuracy. ROC plots the true positive rate ($TPR = \frac{TP}{TP+FN}$) versus the false positive rate ($FPR = \frac{FP}{FP+TN}$) at various threshold settings. A random-guess classifier produces the diagonal line where TPR equals FPR, corresponding to a 50% AUC score. Classifiers with higher AUC scores perform better and the perfect classifier in this setting achieves a 100% AUC score. AUC is a good measure when the dataset has a class imbalance, because AUC is not biased by this imbalance.

NMI: "Normalized Mutual Information" is a technique to evaluate the quality of clustering outcomes using an information theoretical approach [91]. It assumes the availability of a ground truth that represents the correct clusters. Let A be the correct cluster assignment, and suppose that it contains c_A clusters. Let B be the output of a clustering algorithm operating on the same data and producing c_B clusters. We can define a $c_A \times c_B$ confusion matrix N, whose rows correspond to the clusters in A and whose columns represent clusters in B. Each entry N_ij of this confusion matrix reports the number of elements of the correct i-th cluster that happen to be assigned to the j-th cluster by the clustering algorithm. The Normalized Mutual Information is defined as

$$NMI(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i\cdot}N_{\cdot j}}\right)}{\sum_{i=1}^{c_A} N_{i\cdot}\log\left(\frac{N_{i\cdot}}{N}\right) + \sum_{j=1}^{c_B} N_{\cdot j}\log\left(\frac{N_{\cdot j}}{N}\right)}$$

where $N_{i\cdot}$ (resp., $N_{\cdot j}$) is the sum of the elements in the i-th row (resp., j-th column) of the confusion matrix, and N is the sum of all elements of $\mathbf{N}$. The output of this measure is normalized between zero (when the clusters in the two solutions are totally independent), and one (when they exactly coincide). Therefore, the higher the value of NMI, the better the quality of the clusters found by the algorithm. NMI assumes non-overlapping clusters.

A variant of NMI, called *LFK-NMI* after its authors (Lancichinetti, Fortunato, Kertész) is proposed to measure the quality of overlapping clusters [178].

3.4 Limitations of Tools and Data

Many research projects including the ones presented in this dissertation, have limitations and biases introduced by data, methods or tools used for analysis. Some of these limitation can be improved by using more sophisticated techniques, but others have more fundamental roots. Researchers should always be aware of these limitations and interpret their results considering the effects of these limitations. Here I discuss some short-comings of our dataset and techniques.

3.4.1 Twitter Dataset

In this dissertation, we mainly used Twitter data collected from a public stream that corresponds to a random sample of 10% of the public tweets. We also used the Twitter REST API to crawl the most recent tweets produced by certain groups of users.

Limitations and biases of Twitter samples have been studied before [140]. Analysis of different Twitter samples show that the search API over-represents the more central users. Beside the level of activity, researcher should also consider to what extend the population on Twitter represents the actual demographics of the population under study. A study from 2014 shows that the Twitter population is a highly non-uniform sample of the US population [206].

3.4.2 Annotations and Labeled Data

Supervised machine learning algorithms rely on ground-truth data to learn underlying patterns. These labels can be obtained from existing systems or generated by observations.

Human annotation tasks are required in different domains to create ground-truth or training labels. Examples of human-annotation task to create reliable labeled data can be found in building lexicon for word-emotion associations [210], annotating named entities [121], detecting objects on images and videos [264], and many other domains. In this dissertation, we used human annotations to build labeled dataset of human and bot accounts on Twitter.

Crowdsourcing tasks consist of group of a human annotators performing the same or similar tasks. A recent analysis addresses common misconceptions about crowdsourcing tasks [14]. For instance the authors show that disagreement between annotators is not the result of poor quality in the annotation task, but a signal about the difficulty of the task. Agreement between annotators can be used not only to measure the quality of the performed task, but also to identify instances with high disagreement to improve systems.

3.4.3 Methods

When researchers address a new problem, they first consider methods and techniques applicable to the problem. In most of the cases, these methods are the ones that researchers are already familiar with. However, it is always important to know the advantages and limitations of a methodology before implementing it to address a research question.

In this dissertation, we used methods to analyze emotions through off-the-shelf sentiment analysis techniques. Most of these methods rely on lexicons to compute a score for a given text. However, they are not sophisticated enough to identify sarcasm or negation. There exist more sophisticated techniques to learn more nuanced details about the text. Recently, deep learning and vector embedding techniques have become popular. They outperform existing methods, but they have also their own limitations. Vector embeddings for instance, have been shown to be biased on gender due to the nature of the training data [43].

CHAPTER 4

Information Diffusion and Online Discourse

4.1 Diffusion of Trends

Trends represent interesting collective communication phenomena: they are user-generated, continually changing and mostly ungoverned (although orchestrated hijacking attempts have already been observed [52, 240, 241]). So far, trends have been studied as a proxy to detect exogenous real-world events discussed in social media, [5, 23, 87, 246], emerging topics, or news of interest for the online community [60, 183].

But trends are also strongly localized in space and time: the temporal and geographic dimensions play a crucial role to determine the success of a trend in terms of spreading and longevity. We argue that unveiling the spatio-temporal dynamics that drive trending conversations on social media is instrumental to many purposes: from designing successful advertising campaigns, to understanding virality and popularity that characterize some topics. In this work we characterize the relation between trends and geography by tracking and analyzing trending topics on Twitter in 63 main locations of the United States and at the country level, for a period of 50 days in 2013 [116].

In this section we discuss the methodology we followed to generate a dataset of Twitter trends, and the derived temporal dependence network that allows us to unveil the dynamics of trend production and consumption.

Table 4.1: The list of the 63 trend locations in the United States and the relative total number of trends (thousands) they generated in the period between April, 12th and the end of May 2013.

Albuquerque	6.7	Atlanta	5.1	Austin	5.8	Baltimore	5.8
Baton Rouge	6.5	Birmingham	6.1	Boston	5.0	Charlotte	5.2
Chicago	5.2	Cincinnati	5.8	Cleveland	5.4	Colorado Springs	6.7
Columbus	6.0	Dallas-Ft. Worth	5.3	Denver	6.1	Detroit	4.8
El Paso	6.5	Fresno	6.6	Greensboro	5.8	Harrisburg	6.3
Honolulu	6.5	Houston	5.1	Indianapolis	5.9	Jackson	6.8
Jacksonville	6.0	Kansas City	5.7	Las Vegas	5.4	Long Beach	6.5
Los Angeles	5.2	Louisville	5.9	Memphis	6.5	Mesa	6.6
Miami	5.5	Milwaukee	5.8	Minneapolis	5.6	Nashville	6.0
New Haven	5.6	New Orleans	6.2	New York	4.4	Norfolk	6.0
Oklahoma City	5.8	Omaha	6.4	Orlando	5.8	Philadelphia	5.1
Phoenix	5.9	Pittsburgh	5.8	Portland	6.4	Providence	5.9
Raleigh	5.3	Richmond	6.2	Sacramento	5.9	Salt Lake City	6.4
San Antonio	5.8	San Diego	6.2	San Francisco	5.7	San Jose	6.6
Seattle	5.9	St. Louis	5.7	Tallahassee	6.3	Tampa	5.6
Tucson	6.6	Virginia Beach	6.8	Washington	4.7		

4.1.1 Trends Dataset

To build our dataset we monitored in real-time all trends appearing on Twitter for a period of 50 days, starting from April, 12th until the end of May 2013.

The Twitter homepage provides a trends box that contains the top 10 trending hashtags or phrases at any given moment, ranked according to their popularity. Oftentimes, a promoted trend is shown in 1st position — for our analysis we disregarded promoted trends since their popularity is artificially inflated by the advertisement.

Each Twitter user can monitor the trends at the *worldwide*, *country*, or *city* level. Twitter has identified 63 locations in the United States, displayed in Figure 4.4, for which it is possible to follow local trends. The full list of locations is reported in Table 4.1. It is worth noting that some areas are over-represented (for example the East coast and California), while some states (namely, North and South Dakota, Montana, Wyoming, Idaho, and Alaska) are not represented at all.[1]

We deployed a Web crawler to check at regular intervals of 10 minutes the trends of each of these 63 locations and, in addition, those at the country level. We ended up collecting

[1]This has to do with the fact that the activity on Twitter in those states is very low.

11,402 different trends overall: 4,513 hashtags and 6,889 phrases. Table 4.1 also reports how many trends have been observed in each location.

4.1.2 Trend Pathway Backbone Network

To investigate where trends usually start and how they propagate from city to city, we built a temporal dependence network of the 63 locations of the United States represented in our dataset.

This network is directed and weighted: each node corresponds to one of the 63 cities, and the weight of an arc e_{ij} from node i to node j is increased every time location i exhibits a trend before location j. The weight of arc e_{ij} therefore represents the extent to which city i precedes city j in adopting a trend: the higher the weight, the more often location i sets the trends that location j will later adopt.

Due to the fact that the adopted dataset contains a large number of trending hashtags and phrases, the network obtained using the procedure described above is fully-connected. This makes the extraction of relevant connections hard, as each location is connected with all the others and only the weight of the connections vary.

To ease the analysis we applied to this network an edge filtering technique known as multiscale backbone extraction [252]. The goal of this procedure is to retain only those connections that are statistically significant, by removing all edges whose weight does not deviate sufficiently from a null model. The significance level of an edge is determined by a threshold parameter α. Lowering α progressively removes edges and eventually causes the disruption of the network. We tuned α to obtain the backbone network with the minimum number of edges that suffices to maintain all 63 nodes connected ($\alpha = 0.3$). The resulting multiscale backbone of the network is used for the analysis of pathways of trend diffusion, and to investigate trendsetting and trend-following dynamics.

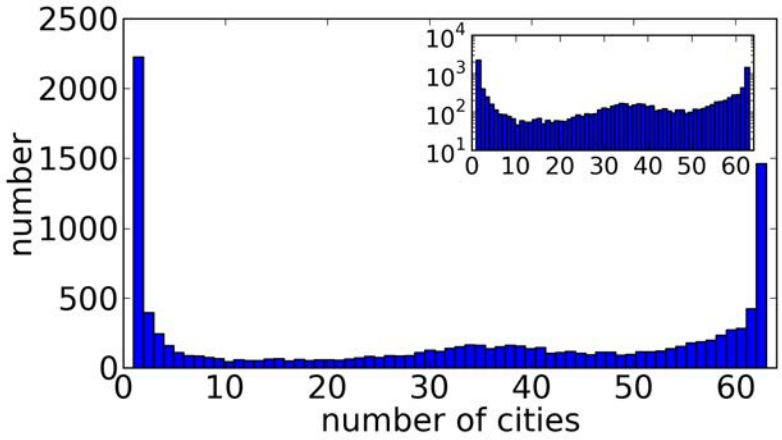

Figure 4.1: Histogram of the number of trends appearing in different number of places. Inset: y-axis reported in a log-scale.

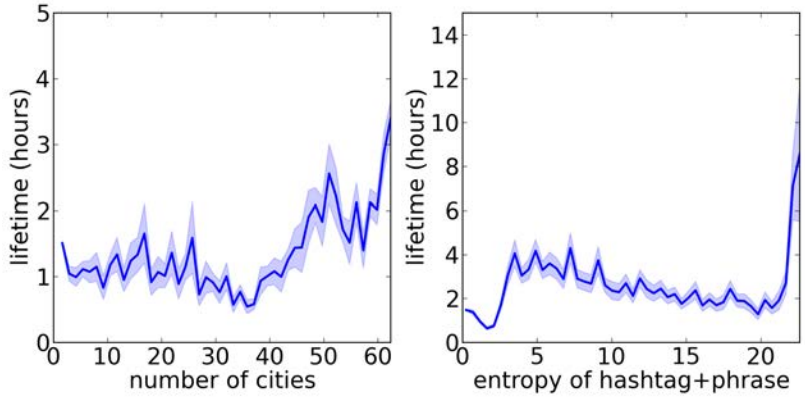

Figure 4.2: Lifetime of a trend. Left: as function of the number of cities in which a trend has appeared. Right: as function of its entropy. In both plots, the dark blue line is the average across trends while the standard error is depicted in light blue.

4.1.3 Spatio-temporal Trend Analysis

In our first experiment we aim to give a statistical characterization of trends: in particular, we start investigating in how many different cities trends appear. In Figure 4.1 we report the number of trends appearing in a given number of distinct locations. Trends follow a bimodal distribution, typically appearing either in one or few locations, or in all or most of them. We can identify three behaviors: (i) a large fraction of trends are localized and not sustained enough to spread from their originating place to others; (ii) another comparably large fraction of trends diffuse all over the cities generating a global phenomenon across

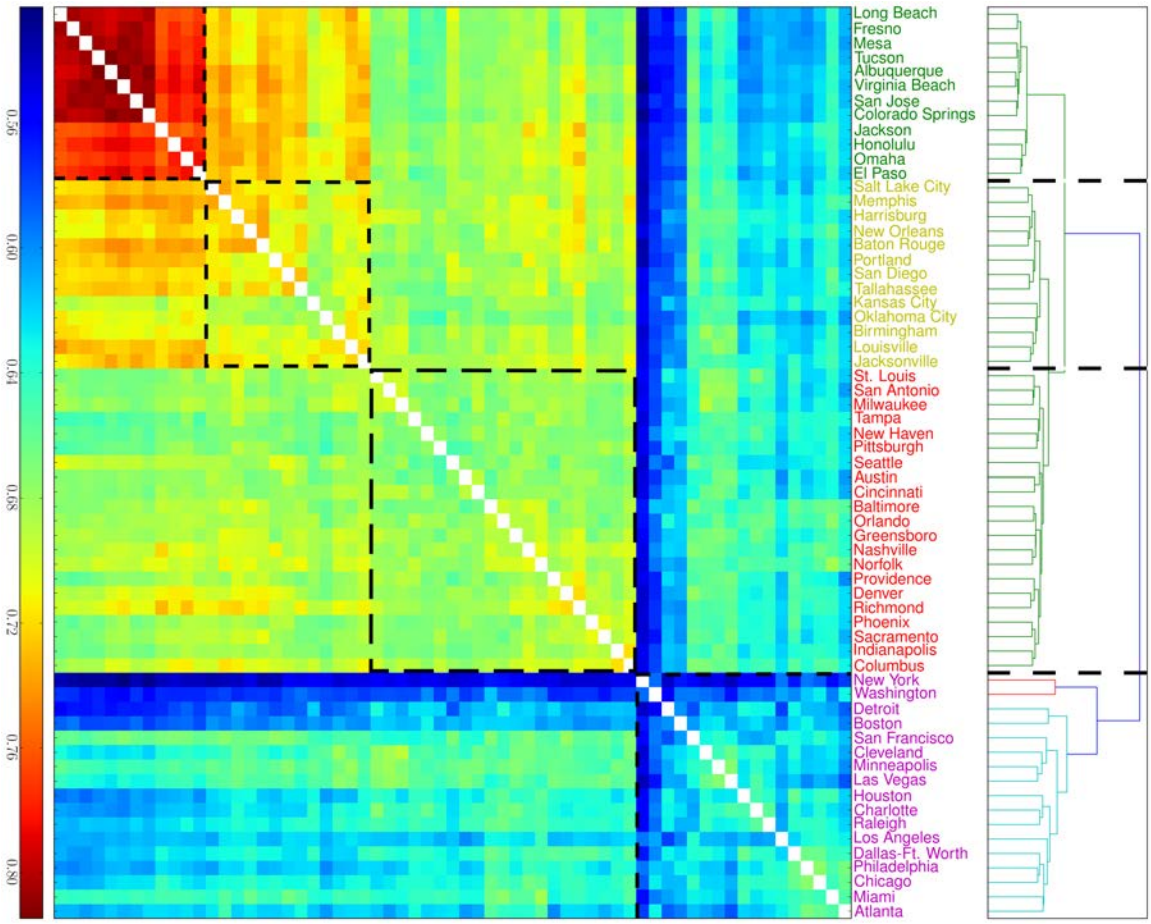

Figure 4.3: Shared trend similarity and hierarchical clustering of the 63 locations.

the country; and (iii) the small remainder diffuse from the originating place to some other places, but fail to achieve global popularity.

The lifetime of trends is broadly distributed: short-lived topics trending for less than 20 minutes amount for more than 68% of the total, and overall trends shorter than six hours cover more than 95% of our sample. Sporadically some trends happen to live a much longer time, with only 0.3% surviving for more than a day.

We now focus on the spatio-temporal dimension of trends, aiming to determine how much time each trend spends in one or several locations. In particular, we calculate the average lifetime of a trend (the average amount of time a given hashtag or phrase is trending somewhere) as a function of the number of cities in which it appears. Figure 4.2 (left panel) reflects the intuition that trends reaching more places live longer.

Another way to determine the relation between the *geographic spread* of trends and their temporal patterns is to measure their lifetime as a function of *entropy*, defined as $\mathcal{S}^j = -\sum_i P_i^j \log P_i^j$, with $P_i^j = \frac{t_i^j}{\sum_k t_k^j}$, where t_i^j is the time topic j has been trending in location i. The entropy is low if the trending topic is concentrated in a few places, and maximal if the topic trends for equal durations of time in all places. Figure 4.2 (right panel) shows that for trends with low entropy (*i.e.*, those concentrated in a single location), the expected lifetime is very short. The lifetime increases significantly (five-fold) for the maximum observed entropy. This analysis reveals a key ingredient for global trend popularity: the trending time of a topic is not only determined by its lifetime in a single location, but also by its geographic spread across many locations.

4.1.4 Geography of Trends

Let us examine the geographic patterns of trends, namely whether geographically close cities share more similar trends than cities that are physically far apart. To determine if this

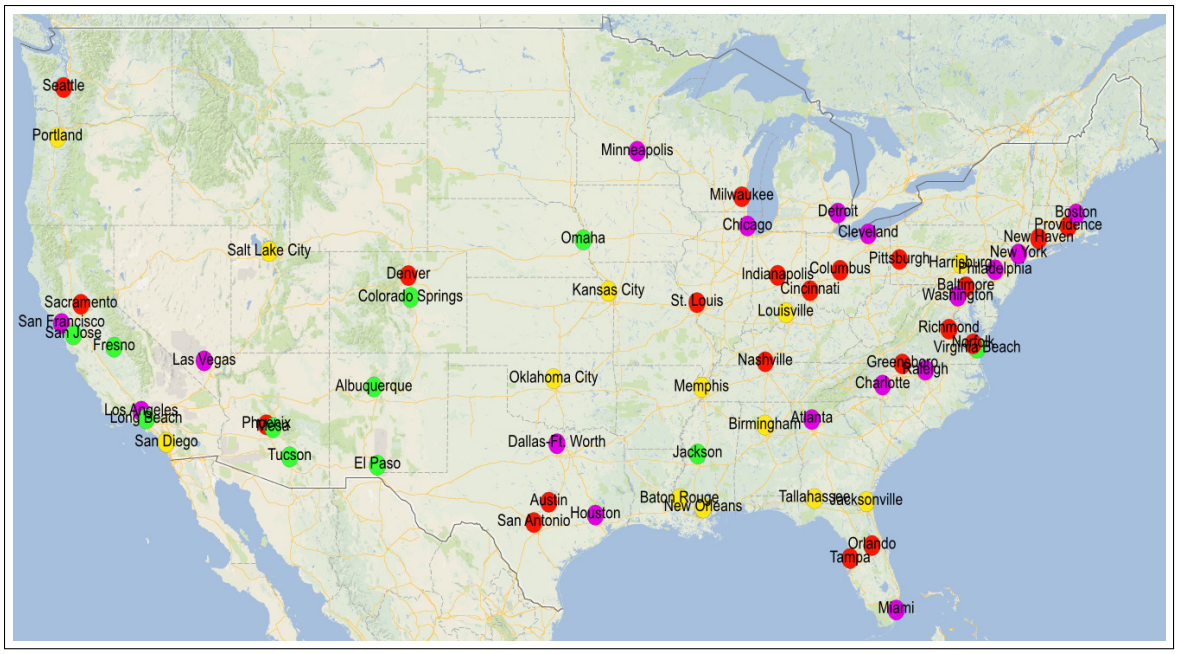

Figure 4.4: geographic representation of the 63 locations and respective clusters.

locality effect exists, we first isolate, for each location i, the set of trends T_i that appeared in that location. Then, for each pair of locations i and j we compute the pairwise Jaccard similarity

$$S_{ij} = \frac{|T_i \bigcap T_j|}{|T_i \bigcup T_j|}. \tag{4.1}$$

The Jaccard similarity ranges between 0 and 1: the higher the value, the more similar the trends exhibited by two different cities. These values of similarity are subsequently passed to a hierarchical clustering algorithm after being transformed in distances: $d_{ij} = 1 - S_{ij}$. This is done to determine whether it is possible to isolate clusters of locations that exhibit similar trends, and, if so, whether these locations are geographically close or spread all over the country. The result is showed in Figure 4.3 and discussed next.

4.1.4.1 Locality Effects

Figure 4.3 is constituted by two parts: a heat-map representing the pairwise Jaccard similarity among locations, and a dendrogram generated according to an agglomerative hierarchical

clustering algorithm using complete linkage. Analyzing the dendrogram we can identify three distinct clusters, whose members (reported in different colors: green, yellow and red) share a high internal similarity in the trends exhibited during the observation period. This cluster emerges applying a cut to the dendrogram for a distance value of 0.5. We can also identify a fourth cluster (in purple, emerging with a dendrogram cut corresponding to a distance value of 0.75) that exhibits a lower internal similarity and whose members show a low similarity with those of other clusters. The four clusters are reported in Table 4.2, and displayed in Figure 4.4.

From the figure we observe that the green, yellow and red clusters are somewhat geographically localized, while the purple one is spread more or less all over the country. In detail, the green cluster, with the highest internal similarity, roughly corresponds to the Southwest of the country. The yellow cluster follows, representing the Midwest and South. The red cluster, which is less localized, matches many locations in the East coast and Midwest. The purple cluster includes several major metropolitan areas [302].

Table 4.2: Clusters of cities according to trend similarity.

Green	Yellow	Red	Purple
Long Beach	Memphis	St. Louis	Washington
Fresno	Salt Lake City	San Antonio	New York
Mesa	Harrisburg	Milwaukee	Detroit
Tucson	New Orleans	Tampa	Boston
Albuquerque	Baton Rouge	Pittsburgh	San Francisco
Virginia Beach	Portland	New Haven	Cleveland
San Jose	Tallahassee	Seattle	Minneapolis
Colorado Springs	San Diego	Cincinnati	Las Vegas
Jackson	Kansas City	Austin	Houston
Honolulu	Oklahoma City	Orlando	Charlotte
El Paso	Birmingham	Baltimore	Raleigh
Omaha	Louisville	Greensboro	Los Angeles
	Jacksonville	Nashville	Dallas-Ft. Worth
		Norfolk	Chicago
		Providence	Philadelphia
		Denver	Miami
		Richmond	Atlanta
		Phoenix	
		Sacramento	
		Columbus	
		Indianapolis	

4.1.4.2 Significance of Geographic Clustering

To determine the statistical significance of the clustering obtained by using the previous method we proceeded as follows: we first computed the distribution of similarity values among all pairs of locations belonging to the same cluster (intra-cluster similarities); then, we did the same for the pairs belonging to different clusters (inter-cluster similarities). After that, we applied a kernel smoothing technique known as Kernel Density Estimation [147] to estimate the probability density functions for our similarity distributions, plotted in Figure 4.5 (the distribution of each cluster is represented by its color corresponding to Table 4.2).

We applied a t-test to determine if any given pair of distributions of intra- and inter-cluster similarity might originate from the same distribution, assessing that all distributions (and, therefore, the clusters) are significant at the 99% confidence level.

We also compared the result of the hierarchical clustering with that of two network clustering algorithms (namely, Infomap [245] and the 'Louvain method' [40]) applied to the trend pathway backbone network. We obtained consistent results in all cases: the only difference was that Seattle was placed in the purple cluster by both network clustering methods.

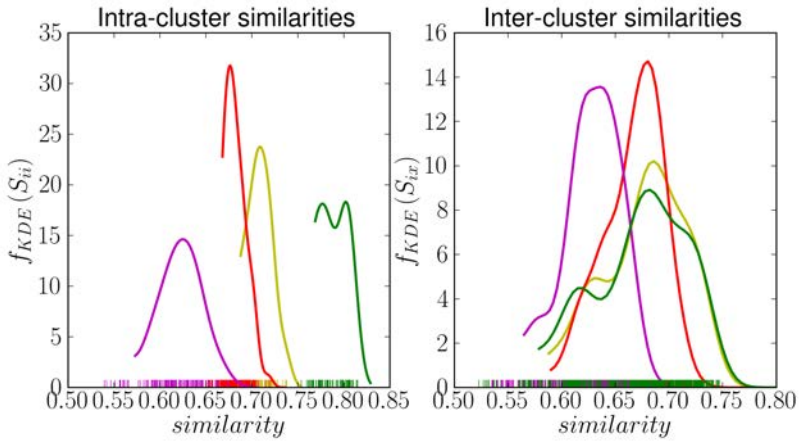

Figure 4.5: Kernel Density Estimation of intra- and inter-cluster similarity of the four clusters.

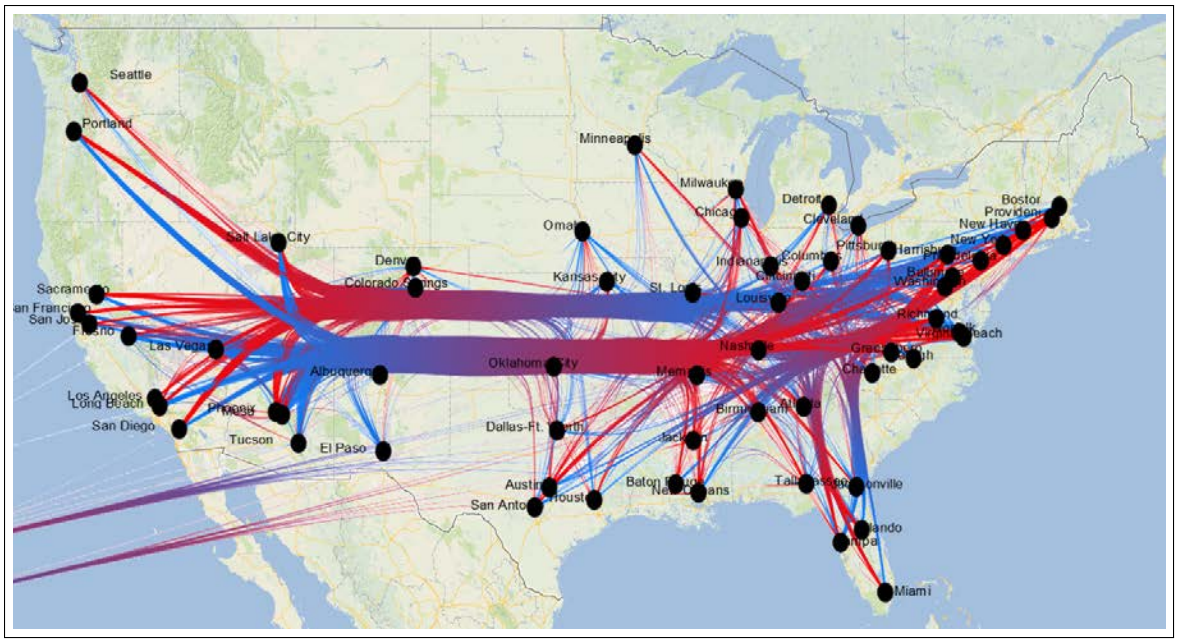

Figure 4.6: Trend pathways in Twitter. Trends spread in the direction from blue to red.

4.1.5 Trend Pathway Analysis

To establish where trends start and what pathways they follow to diffuse in the country, we analyze the multiscale trend pathway backbone network and represented in Figure 4.6 by using a divided edge bundling technique [250]. This visualization strategy has been successfully applied to other geographic networks such as the US airport traffic network (*cf.* [250]). In this node-link representation the edges are bundled taking into account directions and weights. The thicker the bundle, the higher the sum of the weights of connections wrapped in the bundle. In our case, this yields a network visualization that highlights the pathways followed by trends as they flow across the country. In this figure the direction of edges represents the information flow: the tails of the bundles (in blue) show where trends start, the heads of the bundles (in red) point to where the trends arrive. From Figure 4.6 we can draw two observations: first, the presence of a massive backbone that carries the trend flow from the East coast to the West coast and vice-versa. Second, we observe a negligible North-South flow, except for that connecting Florida to the East coast. Moreover,

the fact that the East-to-West flow is well balanced by the that in the opposite direction suggests that we are not simply observing an artifact of the time-zone effect: the West coast contributes to shaping the country trends to a similar extent that the East coast does.

In the backbone network the cities that often generate trends are those with higher fractions of outgoing edges (that is, those that spread their trends to most of the other cities); henceforth we will call them *sources*. Vice-versa, we will call *sinks* those cities with higher fraction of incoming edges. More precisely, since the network we deal with is weighted, we compute the *weighted source-sink ratio* $\omega(n)$ for each node n as

$$\omega(n) = \frac{s_{out}(n)}{s_{in}(n) + s_{out}(n)}, \tag{4.2}$$

where $s_{in}(n)$ (resp., $s_{out}(n)$) is the in-strength (resp., out-strength) of that node. We report in Table 4.3 the top 5 sources and the top 5 sinks of the backbone network. Four out of the five top sources (all but Cincinnati) also happen to be major metropolitan areas. On the other hand, all sinks belong to the Southwest and Midwest parts of the country. Los Angeles and New York (among our top sources) have also been reported in the top 5 hashtag producers worldwide in the recent work by Kamath *et al.* [164].

4.1.6 Trendsetters and Trend-followers

The source-sink analysis presented above triggered our interest in the dynamics of trend popularity. In the following we study trendsetting and trend-following patterns, driven by

Table 4.3: Left: top 5 sources (*i.e.*, trendsetters). Right: top 5 sinks (*i.e.*, trend-followers).

Location	Rank	$\omega(n)$	Location	Rank	$\omega(n)$
Los Angeles	1^{st}	0.806	Oklahoma City	63^{rd}	0.101
Cincinnati	2^{nd}	0.736	Albuquerque	62^{nd}	0.109
Washington	3^{rd}	0.718	El Paso	61^{st}	0.235
Seattle	4^{th}	0.711	Omaha	60^{th}	0.305
New York	5^{th}	0.669	Kansas City	59^{th}	0.352

the following question: *Are trending topics that become popular at the country level produced uniformly by all cities, or preferentially by some of them?*

To answer this question we selected from our dataset all those trends that at some point in time became trending at the country level. This left us with 1,724 hashtags and 2,768 phrases that achieved the highest popularity in the United States, appearing in the top 10 trending topics at the country level. We then selected the set of cities that exhibited each of these trends, and divided them in two categories: those cities in which the hashtag or phrase was trending *before* it became trending at the country level, and those cities that adopted it *after* it became trending at the country level. This allows us to determine what are the cities that contribute more to shaping the trends at the country level, and what are the cities that are more influenced by these global trends: in other words, we can identify trendsetters and trend-followers.

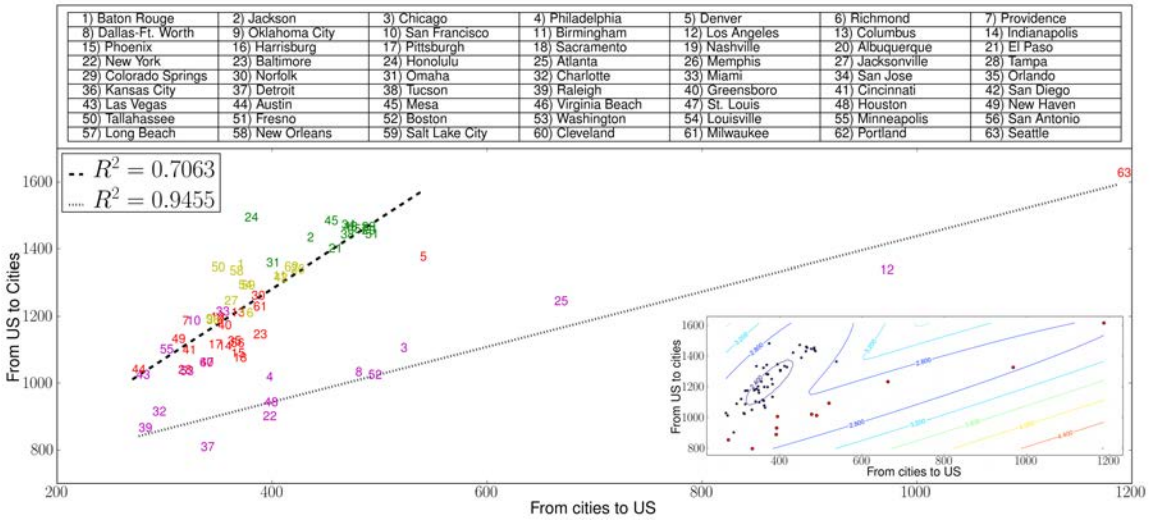

Figure 4.7: Trendsetting vs. trend-following cities. The x-axis shows the number of times a topic trending in a particular city later trends at the country level, while the y-axis shows the number of times of the reverse effect. The inset shows a Gaussian Mixture Model highlighting the two different trendsetting dynamics; the contours represent the standard deviations of each Gaussian distribution. In the main plot, two linear regressions are reported with the corresponding coefficient of determination R^2. City colors correspond to the cluster assignment in Table 4.2.

Figure 4.7 shows the result of this analysis for the hashtags. We can immediately identify

two different classes of cities: the majority of them (*i.e.*, all those in the upper-left part of the main plot) appear to influence country-level trends roughly to the same extent to which they are influenced by the global trends; a second class of cities seem to have a much stronger trendsetting role toward the country.

To assess if these two classes can be significantly distinguished, we use the Expectation Maximization algorithm to learn an optimal Gaussian Mixture Model (GMM); to determine the appropriate number of components of the mixture we perform a 5-fold cross-validation using Bayesian and Akaike information criteria as quality measures, by varying the number of components from 1 to 10. The outcome of the cross-validation determines that the optimal number of components is two, according to both criteria, matching our expectations.

The result of the GMM is showed in the inset of Figure 4.7: each point is assigned to one of the two components yielding two different clusters composed respectively of 11 trendsetting cities (red dots) and 52 trend-following cities (blue stars). The list of trendsetters includes (in ascending order of impact) Raleigh, Detroit, Philadelphia, Houston, New York, Dallas-Ft. Worth, Boston, Denver, Atlanta, Los Angeles, and Seattle. All of them are major metropolitan areas.

To highlight the existence of these two different dynamics we applied a regression analysis approach by fitting two different linear regressions to the points belonging to the classes of trendsetters (coefficient of determination $R^2 = 0.9455$, p-value $p = 3.9 \cdot 10^{-7}$) and trend-followers ($R^2 = 0.7063$, $p < 10^{-10}$). This points out the proportionality that exists between incoming and outgoing trend flows.

We repeated this analysis by making the model even more realistic: for example, we introduced the effect of the time lag, discounting the reward given to those cities that adopt a trend later with respect to the initiators; also, we rewarded only the initiators of each trend, rather than any city that exhibits a given trend before the trending point at the

country level. Making the scenario more realistic did not affect the outcome: in all cases we obtained comparable results.

The fourth, purple cluster identified in clustering analysis deserves further discussion. Differently from the others, this cluster is not geographically well defined (*cf.* Figure **??**) — it contains metropolitan areas spread all over the country. Is the effect of city size sufficient to explain why these metropolitan areas are more influential than others, in the sense that they produce more national trends? It is not obvious that large populations would lead to more national trends: while a larger city produces more tweets and possibly more topic competing for popularity, the number of trends for each city at a given time is bounded to ten, irrespective of the city size. In cities with larger content production, hashtags (or phrases) must appear in more tweets to be listed as a trend, whereas a lower number of tweets is sufficient in cities with smaller content production. As a result, the effect of sheer volume is discounted by construction in the definition of Twitter trends.

Why, then, do the metropolitan areas in the purple cluster play such a trendsetting role? A possible interpretation is offered by noticing the presence in this cluster of some of the major airport hubs of the United States, such as Atlanta, Chicago, and Los Angeles. The list of top US airport hubs [303] is shown in Table 4.4, where we aggregated the traffic by metropolitan area. Surprisingly, 16 out of the 17 locations that constitute the cluster appear in the top 20 air traffic hubs — all of them but Cleveland. On the other hand, some cities in the cluster that do not belong in the top 30 metropolitan areas by population (Charlotte, Raleigh, Las Vegas), do appear among the major air traffic hubs.

The presence of major air traffic hubs among the special class of cities that act as trendsetters suggests an intriguing conjecture, drawing a parallel with the spread of diseases: *Does information travel faster by airplane than over the Internet?* In other words, do conversations and trends spread following social interaction dynamics, like *social butterflies*

Table 4.4: Top 20 cities ranked according to the total volume of flight traffic.

City	Cluster	Rank	Total traffic
New York (JFK, EWR, LGA)	purple	6^{th}, 14^{th}, 20^{th}	54,374,758*
Atlanta (ATL)	purple	1^{st}	45,798,809
Chicago (ORD, MDW)	purple	2^{nd}, 25^{th}	41,603,539*
Miami (MIA, FLL, PBI)	purple	12^{th}, 21^{st}, 54^{th}	33,228,913*
Dallas-Ft. Worth (DFW, DAL)	purple	4^{th}, 45^{th}	31,925,398*
Washington (BWI, IAD, DCA)	purple	22^{nd}, 23^{rd}, 26^{th}	31,431,854*
Los Angeles (LAX)	purple	3^{rd}	31,326,268
Denver (DEN)	red	5^{th}	25,799,832
Charlotte/Raleigh (CLT, RDU)	purple	8^{th}, 37^{th}	24,521,523*
Houston (IAH, HOU)	purple	11^{th}, 32^{nd}	24,082,666*
San Francisco (SFO)	purple	7^{th}	21,284,224
Las Vegas (LAS)	purple	9^{th}	19,941,173
Phoenix (PHX)	red	10^{th}	19,556,189
Orlando (MCO)	red	13^{th}	17,159,425
Seattle (SEA)	red	15^{th}	16,121,123
Minneapolis (MSP)	purple	16^{th}	15,943,751
Detroit (DTW)	purple	17^{th}	15,599,877
Philadelphia (PHL)	purple	18^{th}	14,587,631
Boston (BOS)	purple	19^{th}	14,293,675
Salt Lake City (SLC)	yellow	24^{th}	9,579,836

(*) Sum of the traffic volume of different airports in the same area.

that pass from person to person at the local level, or do they diffuse using traveling people as vectors, similarly to epidemics that take advantage of human mobility [18, 79]?

Further work is needed to explore this conjecture. One possibility would be to measure the correlation between trend overlap among pairs of cities and the corresponding air traffic.

4.2 Spatiotemporal Analysis of Censorship

Many countries want to control online services, if possible, and otherwise apply censorship on content or users. China is one counties that applies strict regulations on social media. The majority of the related research focused on Weibo platform, since the Great Firewall of China prevents foreigner social media services. Here we studied censorship on Twitter and analyzed withheld content. The present work, to the best of our knowledge, is the first to

explore global scale censorship on Twitter.

A well-known example of internet censorship is the Great Firewall of China, which applies broader censorship than any other regulations and has developed its own platforms to control content produced by its citizens. Sina Weibo is the most active social networking site in China and also a replacement for Twitter in the face of China's strict censorship policies. Censorship on the Weibo platform has been studied in terms of identifying topics of censored content, temporal characterization of content deletions, and different censorship practices [19, 315, 316].

Social media platforms are receiving an increasing number of requests for content removal and account closures. When these requests are rejected, some governments simply terminate access to these services. Twitter is one such platform receiving censorship requests and being censored. To respond to censorship requests, Twitter developed a system to withhold tweets and users from particular locations based on the internet protocol (IP) addresses of users. Twitter's approach limits access to content from particular locations as requested by governments while protecting the rights of other users to access content.

This work reports the results of a study exploring the effectiveness of Twitter's censorship policy [281]. As part of our study we also characterized the behavior of users and the effects of censorship on information diffusion. To the best of our knowledge, this paper is the first study exploring censorship applied on Twitter

4.2.1 Twitter Withheld Content

Twitter is a popular micro-blogging platform that is available to millions of people all over the world. Users can interact by creating social ties (friend/follower relations), retweeting content of others to disseminate that content among their friends, and mentioning other users in their posts. Twitter users can post up to 140 characters per tweet including URLs

Figure 4.8: Example of withheld tweet (top) and user (bottom) when they are accessed from censored country.

and external media content, such as pictures and videos, alongside text.

Some of the content shared on Twitter might not be legal under applicable laws in various countries such as copyright, pornography, threatening messages, and insults to other users. Twitter receives requests for removal of content and users from various governments and law enforcement agencies. If removal requests are submitted properly by authorized entities, Twitter grants censorship to these requests.

Another practice applied by Twitter is censoring content by *withholding* it under certain criteria. They can limit access to a particular tweet or user when requested by some country. *Withheld tweets* are censored only by the country that makes such a request to Twitter. Users from countries in "withhold scope" see a notification message about censored tweets in their timelines or in their profiles. Similarly, user accounts can also be withheld under certain conditions and all content created by those users will not be accessible from censoring countries. Examples of notification messages for withheld tweets and users are shown in Fig. 4.8. To apply content censorship, Twitter determines the user locations based on IP addresses. Extensive details about how Twitter processes these requests are found in the Twitter support page.[2]

Twitter also issues quarterly reports with information about the number of requests

[2] `https://support.twitter.com/articles/20169222`

Table 4.5: Descriptions for Twitter censorship decision organized in three categories: withheld, unwithheld, and denied or objected requests. We used explanations from the Twitter transparency pages for each case [273].

Country	Status	Explanation
Turkey	withheld	Court orders directing Twitter to remove content in Turkey regarding violations of personal rights and defamation of both private citizens and/or government officials.
Russia	withheld	Requests received from the Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications (Roskomnadzor) regarding content implicating Federal Law 139 and Federal Law 398. This law allows Russian authorities to restrict access to content that is deemed to be *extremist* or that leads to *mass actions*.
Germany	withheld	Request received from courts and Jugendschutz (child protection) about defamation and usage of prohibited symbols and illegal discriminatory content.
Turkey	unwithheld	Twitter unwithheld content on two separate occasions: when Turkey ban on access to Twitter on March 2014 and Twitters objections to previously censored content were accepted by the courts.
Brazil	unwithheld	Tweets were censored after the request of the Constitutional Court for violating local electoral law and later unwithheld.
Pakistan	unwithheld	Twitter reversed their censorship to content previously withheld due to demand made by the Pakistan Telecommunication Authority for violating local blasphemy law.
Turkey	objected	Twitter filed legal objection with Turkish courts in response to their requests when Twitter believed the order interfered with freedom of expression law or had other deficiencies.
Russia	denied	Twitter denied Russia's requests on silencing popular critics of the Russian government and limiting speech about non-violent demonstrations in Ukraine.

received and summarizes the reasons behind each decision made in a particular country.[3] Examples of these explanations are shown in Table 4.5.

4.2.2 Data Collection

To study censored content on Twitter, we first extracted all withheld tweets and their retweets from our collection (approximately a 10% random sample of all public tweets streamed in real time) starting from June 2013 to December 2015. Twitter API provides meta-data information about withheld tweets. In real-time stream, censorship information

[3] `https://transparency.twitter.com/removal-requests`

Table 4.6: Dataset statistics.

Censored tweets	53,028
Retweets of censored content	99,643
Avg. retweets for censored tweets	64.2
Censored tweet for copyright	310
Unique user created censored content	716
User tweet or retweet censored content	29,619

is not available for the original tweets. We detect censorship through observing retweets of original tweet after the censorship. We identify withheld tweets by monitoring first occurrence of withheld information from the meta-data of retweets. This collection consists of 29,619 unique users who were tweeting and retweeting censored content and 716 of these accounts created at least one censored tweet. We analyzed 2,787 users who tweeted or retweeted at least 3 censored messages and 325 who created at least one censored tweet. We then collected all their tweets, retweets and retweets of their tweets. Basic statistics about our dataset are summarized in Table 4.6.

4.2.3 Censored Tweets

Twitter has been accepting requests for content removals starting from 2012. Since that time several countries have been submitting requests. Consequently, thousands of tweets have been withheld temporarily or permanently. We collected those censored tweets and their retweets. The amount of censored content by countries is shown in Fig. 4.9. We observe consistent statistics with Twitter transparency reports as Turkey (TR), Russia (RU) and Germany (DE) are listed as top countries. We also notice that not all the country codes map a particular geographic location. For instance XZ and XY represent international waters and copyright, respectively.

To characterize the temporal changes of censorship, we study the volume of censored content and their corresponding countries. In Fig. 4.10, we observe a rapid increase in the volume of censored tweets starting from January 2014. The number of censored tweets

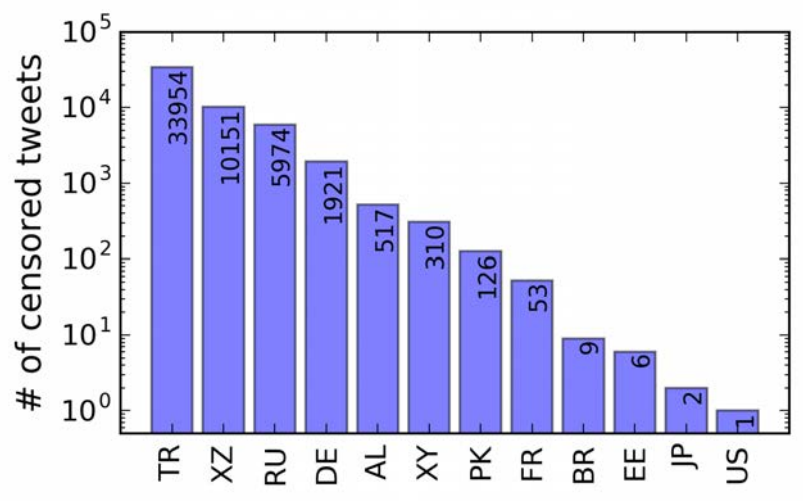

Figure 4.9: Distribution of withheld tweet frequencies by countries

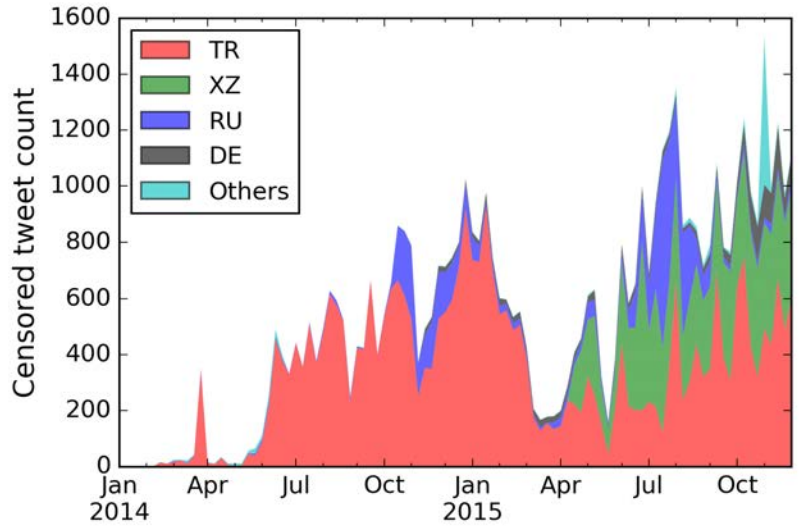

Figure 4.10: Time series of weekly frequencies of withheld content.

reveals an increasing trend. We also observe seasonal changes for the amount of withheld content due to the activity rate of some popular accounts. Active accounts receive more attention as a result of the political discussion that occurred in some countries during that period. For instance Turkey's censorship requests were highest in January 2015 and Russia had more censorship requests during July 2015.

The distribution of censored content per user follows a power law behavior as shown in Fig. 4.11. There are a few users with more than one thousand withheld tweets, but usually

we observe per users 10 to 100 censored tweets. Highly censored users are usually targeted by governments. As a result, users create more content after their first experience of being censored and increase their activity, resulting in more censored content.

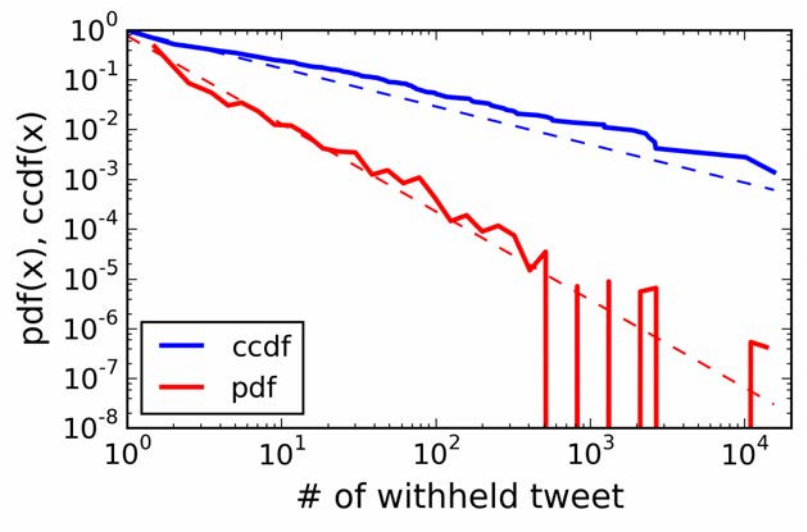

Figure 4.11: Distribution of withheld content per user.

4.2.4 Geographical Censorship

To investigate the effectiveness of IP-based censorship, we analyzed user language and time-zone preferences as proxies for user location. We investigated the relationship between censored countries and time zone and language preferences of retweeting users. In Twitter, users can choose their preferred language, but the language choosen does not necessarily match the language of the content of the tweets. In Figure 4.12, we show co-occurrence between censored country codes and the primary language of users. In this analysis, we can see a relationship between countries and languages due to their political and cultural relevance. For instance the majority of users retweeting censored content in Brazil list their languages as Portuguese, Spanish, and English. Similarly censored content in Russia is retweeted by mostly Ukrainian, Russian, and English-speaking users and content in Turkey is retweeted by Turkish, English, and Arabic speakers. We can also note that English and

Spanish are common languages for users in most of the withheld countries.

An alternative analysis can be carried out using time-zone preferences as a proxy for user location. As we show in Fig. 4.12, users have diverse time-zone preferences regardless of where the content is being censored. For instance in Turkey and Russia, majority of the retweets are created by users from UTC+2 and UTC+3 time zones, which correspond to these countries' local time zones.

These analysis of language and time-zone preferences indicate that the audience of censored content is not bound by geographic location. It is known that citizens of countries with strict internet regulations adopt strategies against censorship by using VPN services or changing DNS settings to access censored sites.

4.3 Roles of Users During Gezi Movement

In this work we focus on the Gezi Park protest, a social uprising whose events unfolded during May and June 2013 in Turkey [224, 284]. Political and policy issues related to this movement have been recently discussed in the social science literature [136, 175]. Here instead we present an empirical analysis of the conversation about Gezi Park that occurred on Twitter. Our goal is to gain insight to the protest discussion dynamics. In particular, we aim at exploring three different aspects of this conversation: *(i)* its spatio-temporal dimension, to determine whether it was concentrated only in the country of inception, or if it acquired significant attention worldwide, and to assess how it started and what trends it generated; *(ii)* what roles individuals played in this conversation and what influence they had on others, and whether such roles changed over time as information was diffused and the protests unfolded; *(iii)* and how the online behavior of individuals changed over time in response to real-world events. To the best of our knowledge, this is the first study to explore the temporal evolution of online user roles and behaviors as a reflection of on-the-ground

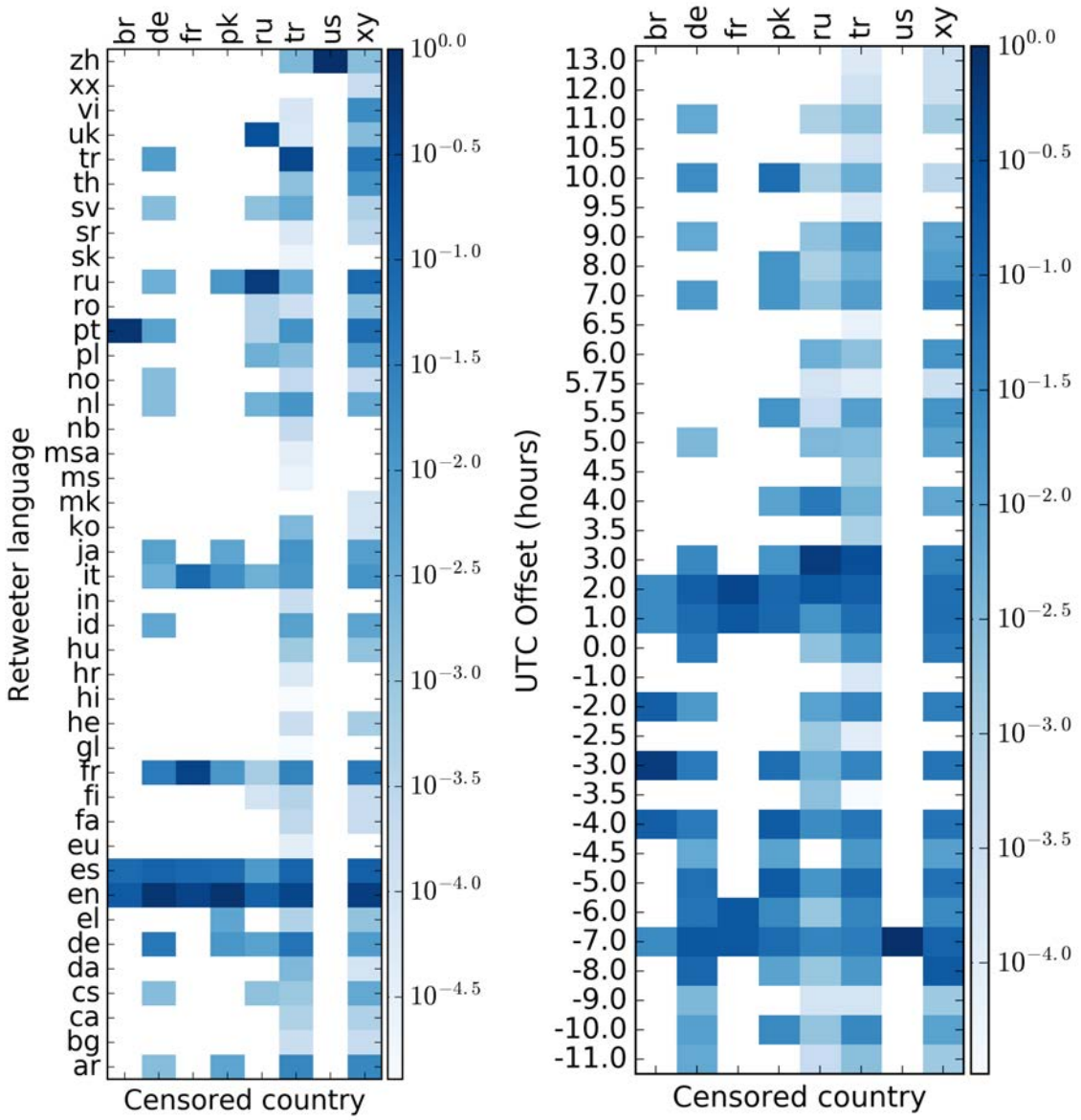

Figure 4.12: Co-occurrence relations for censorship countries (columns) shown for retweeting user's language (left panel) and utf-offset (right panel). Observed frequencies are normalized by shared countries to highlight the distribution of retweeting users.

events during a social upheaval. We do so by means of computational tools and data-driven analyses.

4.3.1 Background of the Protest

In this section we provide some background information about the Gezi Park movement, explaining the context of the protests, the triggers for the mobilization, the timeline of events, and the ways which those events unfolded.

The protests began quietly in an already politically divided Turkey on May 28, 2013 with about 50–100 environmental activists who gathered for a sit-in at Gezi Park in Taksim Square, Istambul. They were there to demonstrate against the destruction of one of the last public green spaces in central Istanbul. The government had slated the space for the construction of a replica of an Ottoman-era barracks that would be the site of luxury residences and a shopping mall. The peaceful encampment successfully resisted the demolition of the park by bulldozers when demonstrators refused to leave. At dawn on the morning of May 30, and then again the next morning, the protesters were attacked by the police using tear gas and water cannons, triggering clashes between authorities and the demonstrators that lasted until the end of the park occupation on June 15. During that time period, the size of the groups of demonstrators escalated to about 10,000 on both the European and Asian sides of the Bosphorus and many thousands more in major cities across the country. The focus of the protests grew from upset over Gezi Park's potential destruction to widespread criticism of the government's increasingly authoritarian practices and intrusions into the private lives of its citizens. As the New York Times reported,

> In full public view, a long struggle over urban spaces is erupting as a broader fight over Turkish identity, where difficult issues of religion, social class and politics intersect. [12]

Throughout the struggle, the protesters, who mostly consisted of middle-class secular Turks but also included some members of left-wing groups and nationalists, used social media to alert others to their plans, urge others to join them, warn participants of police attacks and potential danger spots, provide information about makeshift medical assistance locations,

Table 4.7: List of relevant events during the protest divided in three categories.

	Code	Event date	Event description
Government	A1	2013-05-29	Prime minister Erdogan's statement: "No matter what you do, we took our final decision about Gezi Park."
	A2	2013-06-02	Erdogan refers to protesters as marauders (*çapulcu*).
	A3	2013-06-03	Erdogan says "There is 50 percent, and we can barely keep them at home. But we have called on them to calm down" before his trip to Morocco.
Police	B1	2013-05-30	Police forces raids Gezi Park by using tear gas and destroys tents of protesters without any notice.
	B2	2013-06-03	Official statements about the first death and many injuries all around Turkey.
	B3	2013-06-11	Riot police enters Taksim square with water cannons and uses tear gas against the protesters.
	B4	2013-06-15	Police clears Gezi Park and takes out the protesters. Police starts to stake out Gezi Park.
Protests	C1	2013-06-04	A library is built by the protesters in Gezi Park.
	C2	2013-06-13	Mothers join protests after Huseyin Mutlu's (Governor of Istanbul) calls to mothers to bring their children home.
	C3	2013-06-17	Silent protest in Taksim square held by a standing man. Many others gather after his protest.

and announce their goals. A poll of about 3,000 activists found that the motivation of the demonstrators was their anger with Prime Minister Erdogan and not his political party or his aides. More than 90% of the respondents said they took to the streets because of Erdogan's authoritarian attitude [276].

A detailed timeline of the Gezi Park protests' major events during this period is provided in Table 4.7.

4.3.2 Data Collection

Our analysis is based on data collected from Twitter. Twitter users can post *tweets* up to 140 characters in length, which might contain URLs and media alongside text. Users can also interact with each other through various means, including the creation of directed social links (follower/followee relations), *retweeting* content (*i.e.,* rebroadcasting messages to their followers), and *mentioning* other users in their posts. Tweets may also contain *hashtags*, that are keywords used to give a topical connotation to the tweets (like #direngeziparki and

#occupygezi). Multiple hashtags might co-occur in the same tweet.

The dataset collected for our study comes from a 10% random sample of all tweets streamed in real time, which was stored, post-processed and analyzed in-house. The observation period covers 27 days, from May 25th to June 20th, 2013: this time window started four days prior to the beginning of the Gezi Park events, and fully covered the three weeks during which the main protests unfolded. The short period prior to the protest inception is used as baseline to define user activity and interests.

Our sample not only contains information about the tweets, but also meta-data about the users, including their *screen names*, follower/followee counts, self-reported locations, and more. Additionally, for content posted with a GPS-enabled smartphone, we have access to the geographic location from which the tweets were generated.

To isolate a representative sample of topical discussion about Gezi Park events, we adopted a hashtag seed-expansion procedure [82]: first, we hand-picked the most popular Gezi Park related hashtag (#diregeziparki) and we extracted all tweets containing this hashtag during our 27-day long period of interest. We then built the hashtag co-occurrence list, and we selected the top 100 hashtags co-occurring with our seed (#diregeziparki). We generated our final list of hashtags of interest to include the set of commonly co-occurring hashtags and expanded our dataset collecting all tweets containing any of these hashtags. These hashtags were manually divided in three categories: general-interest hashtags, local protest related ones, and finally those used by government supporters. A detailed list containing the top general-purpose, local-protest, and government-support hashtags are listed in Table 4.8.

Overall, we collected 2,361,335 tweets associated with the Gezi Park movement, generated by 855,616 distinct users and containing a total of 64,668 unique hashtags. Among these 2.3 million tweets, 1,475,494 are retweets and 47,163 are replies from one users to

Table 4.8: Set of hashtags commonly used by protesters and government supporters.

Common hashtags		Local protest hashtags	Gov. supporters' hashtags
#direngeziparki	#bizeheryertaksim	#direnankara	#dunyaliderierdogan
#occupygezi	#gezideyim	#direnbesiktas	#seviyoruzsenierdogan
#eylemvakti	#7den77yedireniyoruz	#direnizmir	#seninleyizerdogan
#occupyturkey	#siddetidurdurun	#direntaksim	#seninleyiztayyiperdogan
#direngezi	#korkakmedya	#direnadana	#youcantstopturkishsuccess
#tayyipistifa	#hukumetistifa	#direndersim	#weareerdogan
#bubirsivildirenis	#dictatorerdogan	#direnistanbul	#yedirmeyiz
#wearegezi		#direnrize	

another. Also, 43,646 tweets have latitude/longitude coordinates. We adopt this subset of geolocated tweets to study the spatio-temporal nature of the protest.

To study type of information carried through conversation and identify roles of participating users during protest, we randomly selected users from our collection. In this work, 135 users and content they created (tweets) and broadcasted (retweets) were extracted for annotation. We annotated 5126 tweets according to rules in our codebook.

Each tweet in our collection was annotated to highlight the message conveyed by the context. Textual information contained in the tweets was studied mainly in three annotation classes:

- *Purpose* categorizes motivation of user for sharing particular content. This annotation highlights motivation behind creating tweets or in broadcasting a particular message.
- *Position* groups different opinions toward particular events or discussions.
- *Information share* classifies type of information conveyed through messages.

In this annotation task, distribution of labels within each category are summarized in Table 4.10. Occurrence of those labels are not homogenous within the categories and some labels are used more frequent than others. If the annotator can not find a match between annotation labels and tweets, they assign those tweets to "others" category.

During the same 27-day long observation period, we monitored the Twitter trends occurring at the country level in Turkey, and at the metropolitan area level in 12 major cities as provided by Twitter, namely: Adana, Bursa, Istanbul, Izmir, Kayseri, Gaziantep, Di-

Table 4.9: Trends in Turkey (country level) and in 12 Turkish cities during the observation period.

Trend Location	Top 5 trending hashtags/phrases
Turkey	Turkey, Necati Şaşmaz, #DirenGeziSeninleyiz, #OyunaGelmiyoruzTakipleşiyoruz, #ProvokatörlereUYMA
Istanbul	Turkey, Necati Şaşmaz, #DirenGeziSeninleyiz, Bruno Alves, #OyunaGelmiyoruzTakipleşiyoruz
Ankara	Turkey, Necati Şaşmaz, Bruno Alves, #DirenGeziSeninleyiz, #ProvokatörlereUYMA
Izmir	Turkey, Necati Şaşmaz, #DirenGeziSeninleyiz, #TatilöncesiTakipleşelim, #ProvokatörlereUYMA
Bursa	Turkey, #TatilöncesiTakipleşelim, Necati Şaşmaz, #KızlarTakipleşiyor, #çapulcularTakipleşirse
Adana	Turkey, #çapulcularTakipleşirse, #TatilöncesiTakipleşelim, Necati Şaşmaz, #DirenGeziSeninleyiz
Gaziantep	Turkey, Necati Şaşmaz, #SesVerTürkiyeBuÜlkeSahipsizDeğil, #DirenGeziSeninleyiz, #OyunaGelmiyoruzTakipleşiyoruz
Konya	#TatilöncesiTakipleşelim, Turkey, #BizimDelilerTakipleşiyor, Necati Şaşmaz, #SesVerTürkiyeBuÜlkeSahipsizDeğil
Antalya	Turkey, #KızlarTakipleşiyor, #CapulchularTakipleşiyor, #TürkiyeBaşbakanınınYanında, Necati Şaşmaz
Diyarbakir	Turkey, Necati Şaşmaz, #DirenGeziSeninleyiz, #OyunaGelmiyoruzTakipleşiyoruz, #ProvokatörlereUYMA
Mersin	#HayranGruplarıTakipleşiyor, Turkey, #TatilöncesiTakipleşelim, #TürkiyemDireniyor, #direnankara
Kayseri	Turkey, Necati Şaşmaz, #DirenGeziSeninleyiz, #Seni_Görünce, #ProvokatörlereUYMA
Eskisehir	Turkey, Necati Şaşmaz, #DirenGeziSeninleyiz, #OyunaGelmiyoruzTakipleşiyoruz, #ProvokatörlereUYMA

yarbakir, Eskisehir, Antalya, Konya and Mersin. The list of top 10 hashtags and phrases trending both at the country level and at the city level were pulled from the platform at regular intervals of 10 minutes. This method [116] is used in our analysis to define the similarity of topical interests and the patterns of collective attention towards Gezi Park conversation in the country. During this period we also monitored worldwide trends to determine if and when the discussion about the protest achieved global visibility. A detailed list of the top popular trending hashtags and phrases for each location and at the country level is provided in Table 4.9.

Annotation category	Category description	# of T	# of T+RT
Purpose	Sharing specific information heard about	849	1559
	Opinion statement	292	506
	General information	289	467
	Links to outside information	230	354
	Support for movement	103	239
	First person witness	130	226
	Ask for help or warn	81	209
	Provide direction	110	184
	Hashtag	86	142
	Information dissemination	62	135
	Media coverage	52	80
	Emotional statement	16	29
Position	General opinion	485	710
	Anger against govt/PM/police	244	435
	Support for movement / motivational	214	363
	Praise or support for groups / individuals	91	183
	Critical statement about people / business or organization of demon	84	161
	Pro-government / police or anti-Gezi opinion	59	80
Info share	Locations of police Tomas, arrests, beatings, info about weapons	414	638
	Scheduled demonstration places, actions of demonstrators	373	596
	Specific info medical, legal, technical, food, safe places	161	297
	Info about specific groups, unions, gays, missing, politicians, etc.	147	215
	About media and availability	103	163

Table 4.10: Descriptions of available annotation categories and their observed frequencies in our dataset. Number of tweets (T) and retweets (RT) for each category excluding "others" category reported.

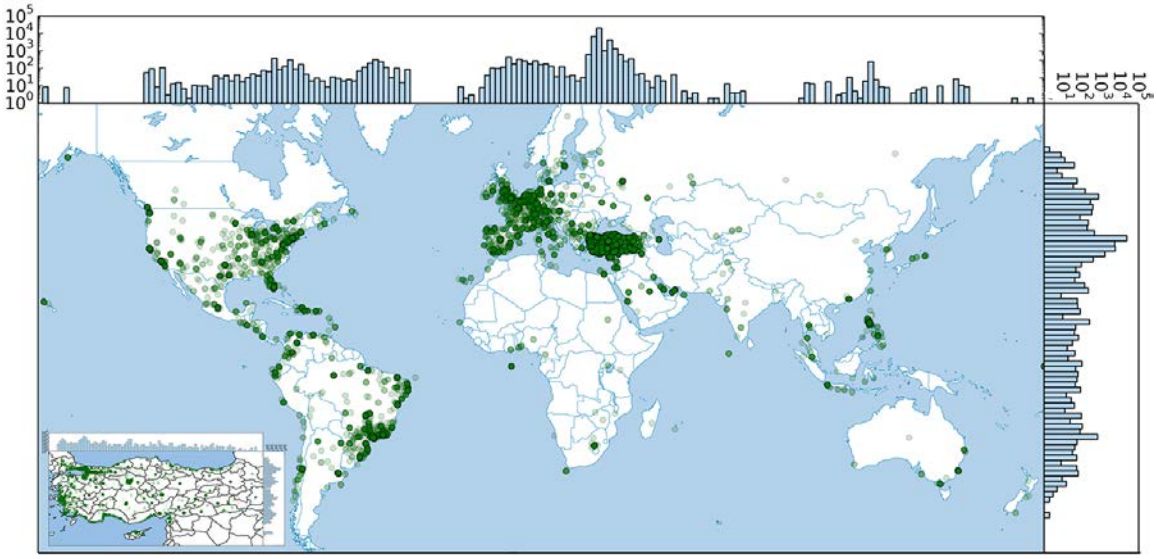

Figure 4.13: Geographic distribution of tweets in our sample related to the discussion of Gezi Park events. The histograms represent the total volume by latitude and longitude. Content production crossed the Turkish national boundaries and spread in Europe, North and South America.

4.3.3 Spatio-temporal Cues of the Conversation

Our first analysis aims at determining the extent to which the discussion about Gezi Park attracted individual attention inside the national boundaries of Turkey, where the movement began, and how much of this conversation spread worldwide.

We focus on the subset of tweets in our dataset that have geo-coordinates attached (in the form of latitude/longitude). Such tweets are likely to be posted by GPS-enabled devices (like smartphones) and represent only a small fraction of total tweets ($\approx$ 1.84% of our sample), which is consistent with similar studies [83]. Yet they provide a very precise picture of the geospatial dynamics of content production. Figure 4.13 maps the sources of these tweets. The figure also shows histograms on the horizontal and vertical axes, that illustrate the distribution of tweets occurring in the corresponding locations, binned by latitude and longitude. From this figure the global nature of the discussion about Gezi Park events clearly emerges. Although a large fraction of tweets originated in Turkey, a significant amount was produced in Europe, North and South America (especially the United States

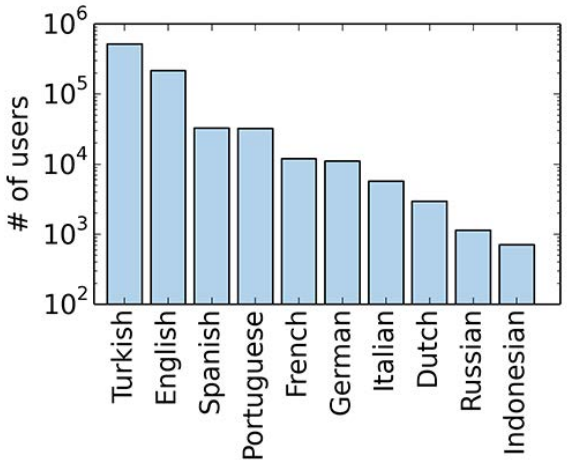

Figure 4.14: Distribution of top 10 languages in tweets about the protest. Language information was extracted from the tweet meta-data.

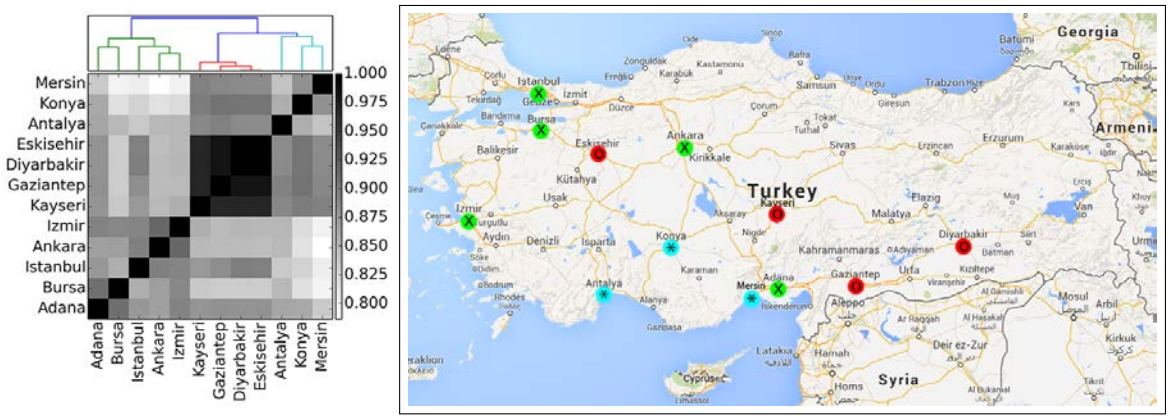

Figure 4.15: Left: Trend similarity matrix for 12 cities in Turkey. From the dendrogram on top we can isolate three distinct clusters. Right: Location of the cities with trend information, labeled by the three clusters induced by trend similarity.

and Brazil). Other noteworthy countries involved in the discussion are the Philippines, Bahrain, Qatar and the United Arab Emirates.

Attention abroad was signaled by the presence of trending hashtags and phrases in the worldwide Twitter trends. Among these, the main protest hashtag, #direngeziparki, trended several times between May 31st and June 2nd, 2013; #TayipIstifa, invoking Erdogan's resignation, appeared on June 6th, 2013. Worldwide attention is also evident in the variety of languages exhibiting hashtags related to the Gezi Park events, as displayed in Figure 4.14. After Turkish, the most popular languages were English, Spanish, and Portuguese.

We also explored the local dimension of the conversation, focusing on the discussion inside the Turkish borders. Our goal was to determine whether any patterns of discussion of similar topics of conversation emerged. In Figure 4.15 we show the trend similarity matrix computed among the sets of trending hashtags and phrases occurring in each of the 12 cities where Twitter trends are monitored. Each location is described by a frequency vector of occurrences of the observed trends. The similarity between pairs of cities is calculated as the cosine similarity of their trends frequency vectors. Above the matrix we show the dendrogram produced by hierarchical clustering, where it is possible to appreciate the separation in three clusters. Such clusters neatly correspond to three different geographic areas of Turkey. Physical proximity seems to play a crucial role in determining the similarity of topical interests of individuals, consistent with other recent results [116].

The clusters found with our trend similarity analysis also seem to match the Turkish geopolitical profiles. Eskisehir, Kayseri and Gaziantep (in the red cluster) are all central Anatolian cities where the president's party (AKP) has a stronghold (though the CHP opposition party edged out the AKP in the March 2014 mayoral race); they are more culturally conservative and homogeneous. Izmir, Istanbul, Bursa, Ankara, and Adana (green cluster) are the largest cities in Turkey with diverse populations. Finally, Antalya and Mersin (blue cluster) are seacoast cities that are known for supporting the one of the main opposition parties (CHP or MHP). Further work is needed to understand why Konya is assigned to this cluster, as it is considered a major religiously conservative center (where the AKP mayoral candidate secured more than 64% of the vote in the 2014 mayoral elections) that has little in common with the Mediterranean cities.

Let us explore the temporal dimension of the Gezi Park discussion. We wanted to determine whether the activity on social media mirrored on-the-ground events, and whether bursts of online attention coincided with real-world protest actions. We analyzed the time

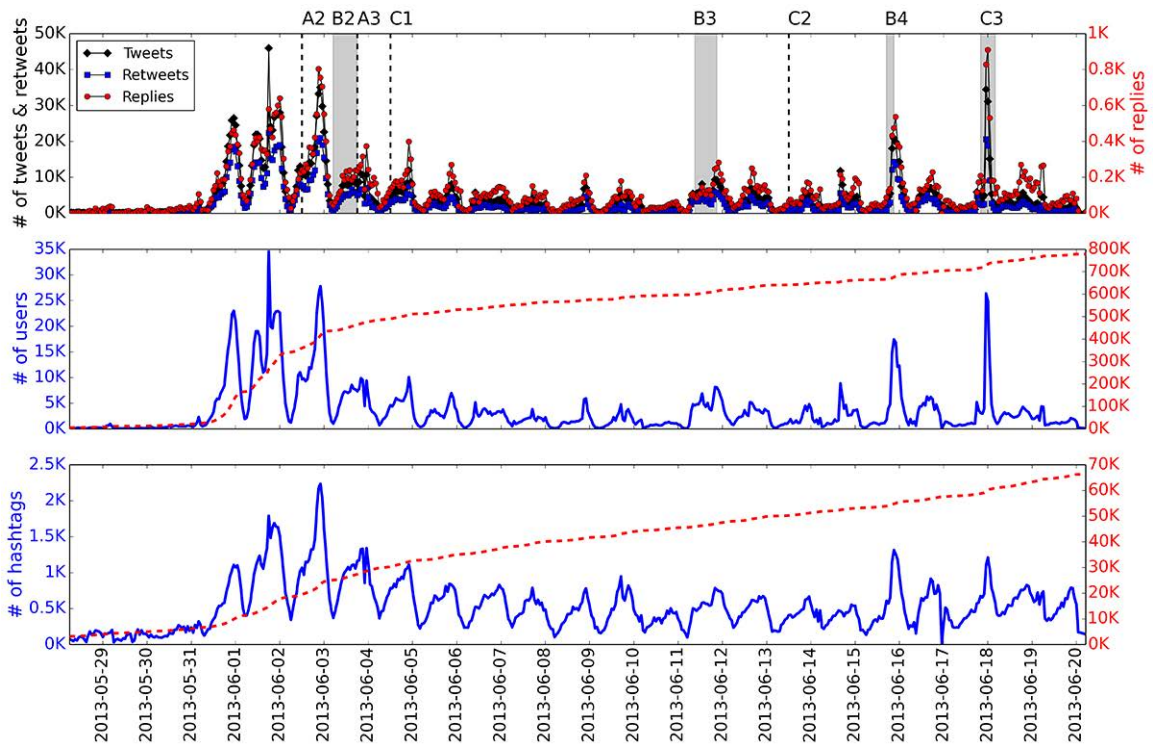

Figure 4.16: Hourly volume of tweets, retweets and replies between May 30th and June 20th, 2013 (top). The timeline is annotated with events from Table 4.7. User (center) and hashtag (bottom) hourly and cumulative volume of tweets over time.

series of the volume of tweets, retweets and replies occurring during the 27-day-long observation window, as reported in Figure 4.16 (top panel). The discussion was driven by bursts of attention that largely corresponded to major on-the-ground events (cf. Table 4.7), similar to what has been observed during other social protests [84]. It is also worth noting that the numbers of tweets and retweets are comparable throughout the entire duration of the conversation, suggesting a balance between content production (*i.e.,* writing novel posts) and consumption (*i.e.,* reading and rebroadcasting posts via retweets). In the middle panel of Figure 4.16 we report the number of users involved in the conversation at a given time, and the cumulative number of distinct users over time (dashed red line); similarly, in the bottom panel of the figure, we show the total number of hashtags related to Gezi Park observed at a given time, and the cumulative number of distinct hashtags over time. We note that approximately 60% of all users observed during the entire discussion joined in the very first

few days, whereas additional hashtags emerged at a more regular pace throughout a longer period. This suggests that the conversation acquired traction immediately, and exploded when the first on-the-ground events and police action occurred.

4.3.4 User Roles and Their Evolution

Our second experiment aims at investigating what roles users played in the Gezi Park conversation and how they exercised their influence on others. We also seek to understand whether such roles changed over time, and, if so, to what extent such transformation reshaped the conversation.

Figure 4.17 shows the distribution of social ties reporting the two modalities of user connectivity, namely followers (incoming) and followees (outgoing) relations. The dark cells along the diagonal indicate that most users have a balanced ratio of ingoing and outgoing ties. Users below the diagonal follow more than they are followed. Note that most users are allowed to follow at most 1000 people. Finally, above the diagonal, we observe users with many followers. Note the presence of extremely popular users with hundreds of thousands or even millions of followers. The number of followers has a broad distributions and seems largely independent of the number of followees.

The presence of highly followed users in this conversation raises the question of whether their content is highly influential. Following a methodology inspired by González-Bailón *et al.* [138], we determined user roles as a function of their social connectivity and interactions. Figure 4.18 gives an aggregated picture of the distribution of user roles during the Gezi Park conversation. The y-axis shows the ratio between number of followees and followers of a given user; the x-axis shows the ratio between the number of retweets produced by a user and the number of times other users retweet that user. In other words, the vertical dimension represents social connectivity, whereas the horizontal dimension accounts for information

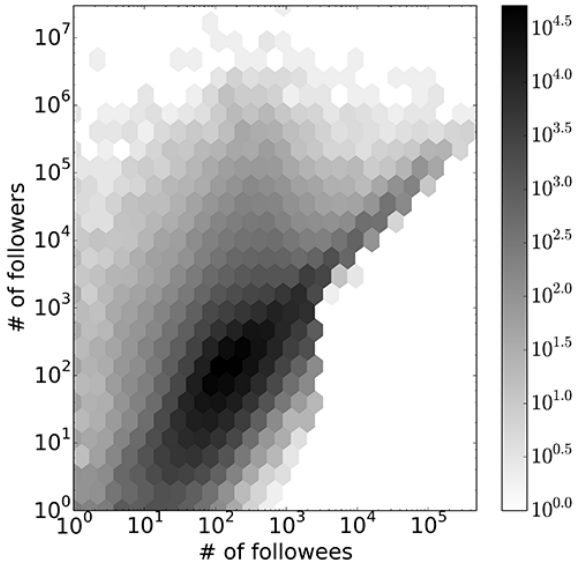

Figure 4.17: Distribution of friends and followers of users involved in the Gezi Park conversation.

diffusion. We can draw a vertical line to separate influential users on the left (*i.e.,* those whose content is most often retweeted by others) and information consumers on the right (those who mostly retweet other people's content). Influential users can be further divided in two classes: those with more followers than followees (bottom-left) and those with fewer followers (top-left), which we call *hidden influentials*. Similarly, information consumers can be divided in two groups–rebroadcasters with a large audience (bottom-right), and common users (top-right).

Figure 4.18 shows a static picture of aggregated data over the 27-day observation period. To study how roles evolve as events unfold, we carried out a longitudinal analysis whose results are provided in Figure 4.19. This figure shows the average displacement of each role class, and the number of individuals in each class (circles), for each day. The displacement is computed in the role space (that is, the space defined by the two dimensions of Figure 4.18). Larger displacements suggest that individuals in a class, on average, are moving toward other roles.

Various insights emerge from Figure 4.19: first, we observed that the classes of informa-

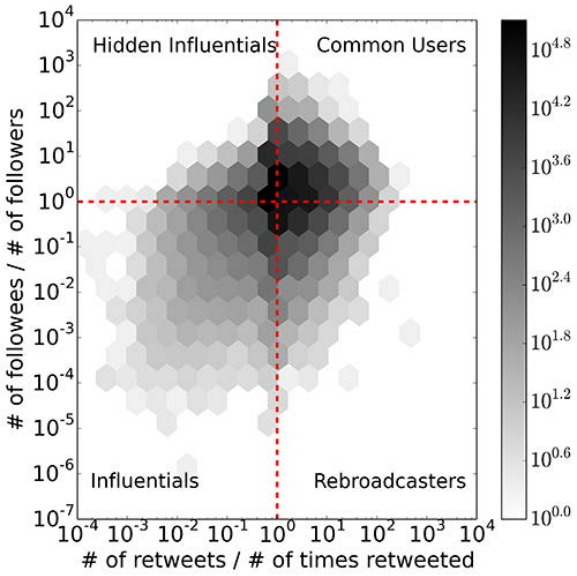

Figure 4.18: Distribution of user roles as function of social ties and interactions.

tion producers (influentials and hidden influentials) are relatively stable over time; together they include more than 50% of users every day, suggesting that many individuals in the conversation had large audiences, and the content they produced was heavily rebroadcasted by others (information consumers as well as other influentials). On the other hand, information consumers show strong fluctuation: starting from an initial configuration with stable roles (May 29–31), common users and rebroadcasters subsequently exhibit large aggregate displacements in the role space (June 1–4). We also note a redistribution of the users in each role: at the beginning of the protest a large fraction represents common users and rebroadcasters, while, as time passed and events unfolded, these two classes shrank. This suggests that common users and rebroadcasters acquired visibility and influence over time: some fraction of these users moved from the role of information consumers to that of influentials, such that their content wass consumed and rebroadcasted by others. In other words, the discussion became more *democratic* over time, in that the control of information production was redistributed to a larger population, and individuals acquired influence as the protests unfolded.

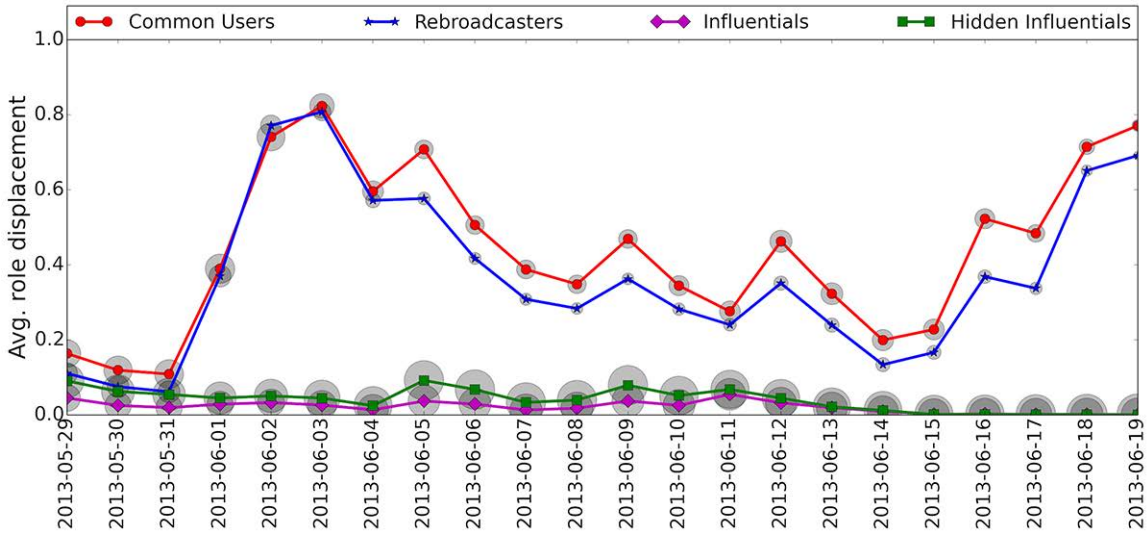

Figure 4.19: Average displacement of roles over time for the four different classes of roles. The size of the circles represents the number of individuals in each role.

4.3.5 Clustering User Roles Using Annotated Data

Clustering users based on their social connectivity and production of content provides overall view about dynamics and behavior. Using information obtained from annotated content, we can further identify users with different roles. In this section, we are using unsupervised clustering framework to identify groups of users that produce or broadcast similar contents according to our annotated dataset.

We can study individuals by their content production and consumption preferences. Individuals share various content based on their motives about participating the protests. Analyzing those content helps us to obtain crucial information about users. We compared users by annotations of their tweets. Each user is represented as a vector of tweet annotations where each category and their observed frequencies represent a feature. We also consider tweets and retweets separately to highlight differences between information need and creation. To compare users, we compute cosine similarity of their vector representations.

We considered 3 different annotations of tweets, namely “purpose”, “position”, and “information share”. Combinations of those 3 type of annotation is also considered.

We clustered users by annotations of their contents using hierarchical clustering. In this technique, distance between each pair of items used to compute clusters from bottom up by agglomerating similar users in each step. In this analysis, we used complete linkage to merge clusters in the hierarchy. One of the advantages of using hierarchical clustering is to tuning threshold to decide number of clusters. We explored different threshold values to observe outcomes of clustering algorithm. In Figure 4.20, we presented clusters of users in distance matrix and hierarchy of clusters. In this analysis annotations on information share, purpose and position were used together.

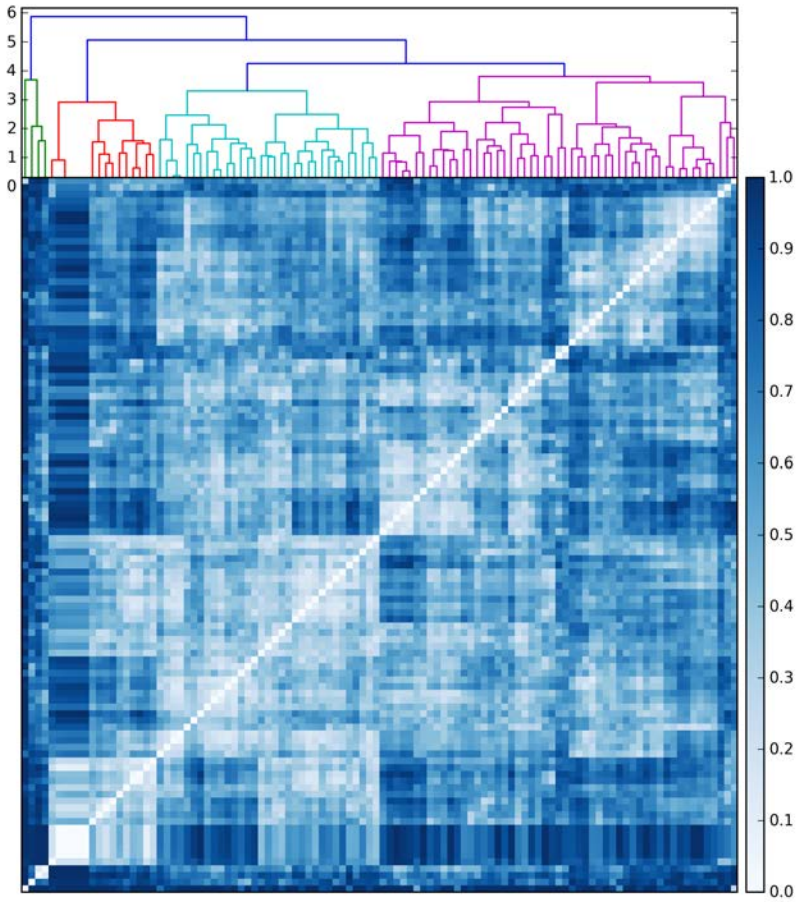

Figure 4.20: Hierarchical clustering of the users by using their similarities based on content annotations.

Once we explore alternative clustering outcomes, we can select most appropriate clustering by looking distance matrix and dendrogram in hierarchical clustering shown in Fig.6. In

this representation, we can choose 5 clusters as represented by different colors and branches in dendrogram.

We can further explore content created or shared by average users in each group. We can observe differences between average behaviors of groups. In Table 4.11, we summarized most common annotation types for each group. For instance group 1 has only 4 users and represent smallest group among 5 clusters. In this small group, we observe tweets related to witnessing and information related to on-the-ground events. Users in this group seem very active for protests. Another dimensions that we can investigate differences between groups are annotation category and message type. Groups 2 and 4 mostly create their original content, but 3rd and 5th groups tend to broadcast content produced by others. Content from information share category, which aim to inform others about scheduled events and locations of police mostly created by group 4 and broadcasted by group 5. Groups 2 and 3 are explicitly announcing their position and opinions about protests. All these groups produce or broadcast particular type of content. Some of those contents overlap between groups, but amount of contents differ.

4.3.6 Online Behavior and Exogenous Factors

Our concluding analysis focused on the way on-the-ground events affected online user behavior. While analyzing our dataset we noticed an abnormal number of screen name changes, as reported in Figure 4.21 (the screen name, not to be confused with the user name, is the name displayed in one's Twitter account). Many users changed their screen names five or more times. This was an unusual observation that attracted our attention.

Further investigation revealed a collective synchronization process, as displayed in Figure 4.22. The changes in screen names represent reactions of users involved in the Gezi Park conversation to external events: these users changed their Twitter screen names to

ID (Size)	Category	Description	Value
C1 (4)	purpose	First person witness	11.75 (T)
	info_share	Locations of Police Tomas, Arrests, Beatings, Info about Weapons	7.0 (T)
	info_share	Scheduled Demon Places, Actions of Demonstrators	5.75 (T)
	purpose	Hashtag	4.0 (T)
	purpose	Links to outside information	1.5 (T)
C2 (16)	purpose	Others	0.75 (T)
	position	General Opinion	0.38 (T)
	purpose	Sharing specific information heard about	0.31 (T)
	info_share	Locations of Police Tomas, Arrests, Beatings, Info about Weapons	0.25 (T)
	position	General Opinion	0.25 (R)
C3 (33)	position	General Opinion	0.18 (T)
	purpose	Sharing specific information heard about	0.18 (T)
	purpose	Others	0.15 (R)
	purpose	Sharing specific information heard about	0.15 (R)
	info_share	Scheduled Demon Places, Actions of Demonstrators	0.12 (T)
C4 (28)	purpose	Sharing specific information heard about	0.32 (T)
	info_share	Scheduled Demon Places, Actions of Demonstrators	0.18 (T)
	info_share	About Media and availability	0.14 (T)
	purpose	Links to outside information	0.14 (T)
	info_share	Locations of Police Tomas, Arrests, Beatings, Info about Weapons	0.11 (T)
C5 (25)	purpose	Opinion statement	0.32 (T)
	purpose	Sharing specific information heard about	0.32 (R)
	purpose	General information	0.28 (R)
	info_share	Locations of Police Tomas, Arrests, Beatings, Info about Weapons	0.2 (R)
	purpose	Support for movement	0.2 (T)

Table 4.11: Average behavior of users in each cluster. Most common 5 activity reported for each group along with their amount and type of share (being retweet or tweet)

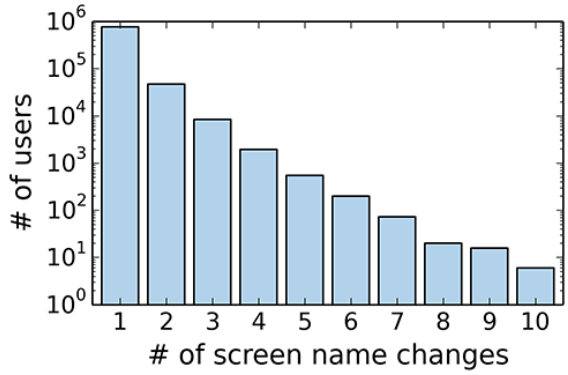

Figure 4.21: Distribution of the number of screen name changes among users during the Gezi Park events.

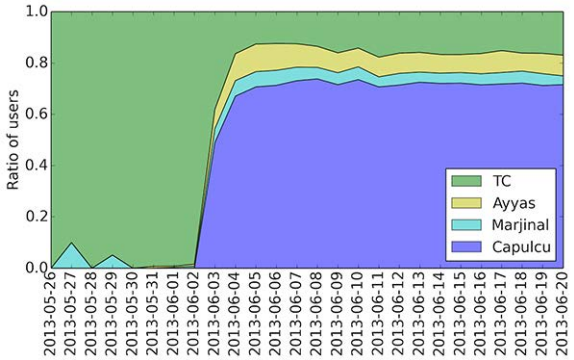

Figure 4.22: Among the many users who changed screen names, this chart plots the fractions who adopted different nicknames over time in respons to external events.

reflect sobriquets attributed to them by their political leaders. One example is the adoption of "TC" (standing for *Turkiye Cumhuriyeti* — Turkish Republic). As a reaction to identity issues, several users started using TC in front of their screen names. Another relevant example is Erdogan's speech of June 2, during which he referred to protesters as marauders (*çapulcu*), marginals (*marjinal* or drunks (*ayyas*). Individuals responded by changing their screen names to include such nicknames as a sign of protest against the government's attempt to discredit the protest participants and minimize the relevance of their actions. This phenomenon illustrates how online and offline worlds are tightly interconnected, deeply affecting each other.

CHAPTER 5

Early Detection of Promoted Campaigns

An increasing number of people rely, at least in part, on information shared on social media to form opinions and make choices on issues related to lifestyle, politics, health, and products purchases [17, 44, 226]. Such reliance provides a variety of entities — from single users to corporations, interest groups, and governments — with motivation to influence collective opinions through active participation in online conversations. There are also obvious incentives for the adoption of covert methods that enhance both perceived and actual popularity of promoted information. There are abundant recently reported examples of abuse: astroturf in political campaigns, or attempts to spread fake news through social bots under the pretense of grassroots conversations [35, 115, 240]; pervasive spreading of unsubstantiated rumors and conspiracy theories [34]; orchestrated boosting of perceived consensus on relevant social issues performed by governments [255]; propaganda and recruitment by terrorist organizations, like ISIS [31, 118]; and actions involving social media and stock market manipulation [277].

The situation is ripe with dangers as people are rarely equipped to recognize propaganda or promotional campaigns as such. It can be difficult to establish the origin of a piece of news, the reputation of its source, and the entity behind its promotion on social media, due both to the intrinsic mechanisms of sharing and to the high volume of information that competes for our attention. Even when the intentions of the promoter are benign, we easily

interpret large (but possibly artificially enhanced) popularity as widespread endorsement of, or trust in, the promoted information.

There are at least three questions about information campaigns that present scientific challenges: what, how, and who. The first concerns the subtle notion of trustworthiness of information, ranging from verified facts [74], to rumors and exaggerated, biased, unverified or fabricated news [34, 240, 314]. The second considers the tools employed for the propaganda. Again, the spectrum is wide: from a known brand that openly promotes its products by targeting users that have shown interest, to the adoption of social bots, trolls and fake or manipulated accounts that pose as humans [76, 115, 148, 282]. The third question relates to the (possibly concealed) entities behind the promotion efforts and the transparency of their goals. Even before these question can be explored, one would need to be able to *identify* an information campaign in social media. But discriminating such campaigns from grassroots conversations poses both theoretical and practical challenges. Even the very definition of "campaign" is conceptually difficult, as it entangles the nature of the content (e.g., product or news), purpose of the source (e.g., deception, recruiting), strategies of dissemination (e.g., promotion or orchestration), different dynamics of user engagement (e.g., the aforementioned social bots), and so on.

This paper takes a first step toward the development of computational methods for the *early detection* of information campaigns. In particular, we focus on trending memes and on a special case of promotion, namely advertisement, because they provide convenient operational definitions of social media campaigns. We formally define the task of discriminating between organic and promoted trending memes. Future efforts will aim at extending this framework to other types of information campaign.

Table 5.1: Summary statistics of collected data about promoted and organic trends on Twitter.

	Promoted		Organic	
Dates	1 Jan– 31 Apr 2013		1–15 Mar 2013	
No. trends	75		852	
	mean	st. dev.	mean	st. dev.
Avg. no. tweets	2,385	6,138	3,692	9,720
Avg. no. unique users	2,090	5,050	2,828	8,240
Avg. retweet ratio	42%	13.8%	33%	18.6%
Avg. reply ratio	7.5%	7.8%	20%	21.8%
Avg. no. urls	0.25	0.176	0.15	0.149
Avg. no. hashtags	1.7	0.33	1.7	0.78
Avg. no. mentions	0.8	0.28	0.9	0.35
Avg. no. words	13.5	2.21	12.2	2.74

5.1 The Challenge of Identifying Promoted Content

On Twitter, it is common to observe *hashtags* — keywords preceded by the # sign that identify messages about a specific topic — enjoying sudden bursts in activity volume due to intense posting by many users with an interest in the topic [181, 215, 310]. Such hashtags are labeled as *trending* and are highlighted on the Twitter platform. Twitter algorithmically identifies trending topics in a predetermined set of geographical locations. Although Twitter recently included personalized and clustered trends, the ones in the collection analyzed here correspond to single hashtags selected on the basis of their popularity. Unfortunately, detailed knowledge about the algorithm and criteria used to identify organic trends is not publicly available [274]. Other hashtags are exposed prominently after the payment of a fee by parties that have an interest in enhancing their popularity. Such hashtags are called *promoted* and often enjoy subsequent bursts of popularity similar to those of trending hashtags, therefore being listed among trending topics.

Of course, once Twitter labels a hashtag as trending, it is not necessary to detect whether or not it is promoted — this information is disclosed by Twitter. However, since it is difficult to manually annotate a sufficiently large datasets of campaigns, we use organic and promoted trending topics as a *proxy* for a broader set of campaigns, where promotion mechanisms may be hidden. Our data collection methodology provide us with a large

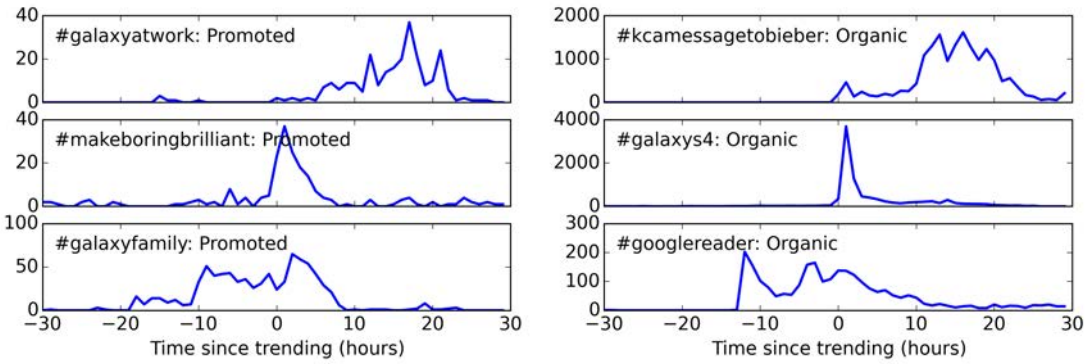

Figure 5.1: Time series of trending hashtags. Comparison of the time series of the volume (number of tweets per hour in our sample) relative to promoted (left) and organic (right) trends with similar temporal dynamics.

source of reliable "ground truth" labels about promotion, which represent an ideal testbed to evaluate detection algorithms. These algorithms have to determine whether or not a hashtag is promoted based on information that would be available even in cases where the nature of a trend is unknown. We stress that our goal of distinguishing mechanisms for promoting popular content is different from that of predicting viral topics, an interesting area of research in its own right [65, 66, 297].

Discriminating between promoted and organically trending topics is not trivial, as Table 5.1 illustrates — promoted and organic trending hashtags often have similar characteristics. One might assume that promoted trends display volume patterns characteristic of exogenous influence, with sudden bursts of activity, whereas organic trends would conform to more gradual volume growth patterns typical of endogenous processes [181, 216, 263]. However, Fig. 5.1 shows that promoted and organic trends exhibit similar volume patterns over time. Furthermore, promoted hashtags may preexist the moment in which they are given the promoted status and may have originated in an entirely grassroots fashion. It is therefore conceivable for such hashtags to display features that are largely indistinguishable from those of other grassroots hashtags about the same topic, at least until the moment of promotion.

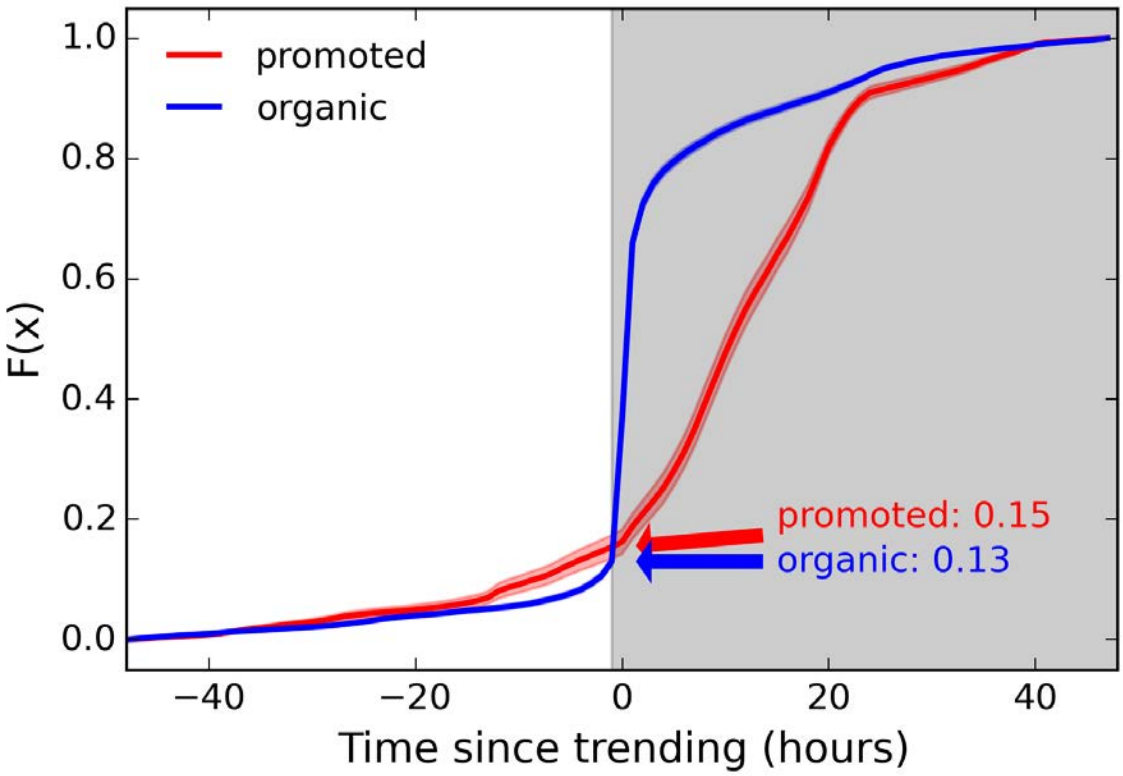

Figure 5.2: Cumulative fraction of tweets as a function of time. On average, only 13% of the tweets in the organic class and 15% of the tweets in the promoted class are produced prior to the trending point. The majority of tweets are observed after the trending point, with a rapid increase around trending time.

The analysis in this paper is motivated by the goal of identifying promoted campaigns at the earliest possible time. The early detection task addresses the difficulty of judging the nature of a hashtag using only the limited data available immediately before trending. Fig. 5.2 illustrates the shortage of information available for early detection. It is also conceivable that once the promotion has triggered interest in a hashtag, the conversation is sustained by the same mechanisms that characterize organic diffusion. Such noise around popular conversations may present an added difficulty for the early detection task.

The major contribution of this paper, beyond formulating the problem of detection of campaigns in social media, is the development and validation of a supervised machine learning framework that takes into consideration the temporal sequence of messages associated

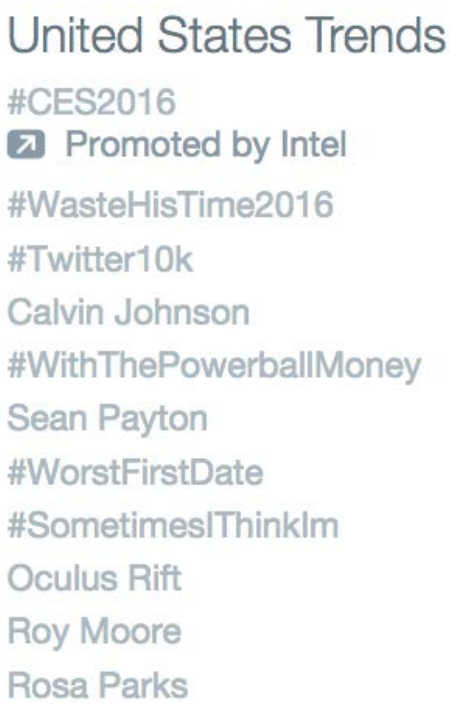

Figure 5.3: Screenshot of Twitter U.S. trends taken on Jan. 6, 2016. The hashtag `#CES2016` was promoted on this date.

with a trending hashtag on Twitter and successfully classifies it as either "promoted" (advertised) or "organic" (grassroots). The proposed framework adopts time-varying features built from network structure and diffusion patterns, language, content and sentiment information, timing signals, and user meta-data. In the following sections we discuss the data we collected and employed, the procedure for feature extraction and selection, the implementation of the learning framework, and the evaluation of our system.

5.2 Data and Methods

5.2.1 Dataset Description

The dataset adopted in this study consists of Twitter posts (*tweets*) that contain a trending hashtag and appeared during a defined observation period. Twitter provides an interface that lists trending topics, with clearly labeled *promoted* trends at the top (Fig. 5.3). We

crawled the Twitter webpage at regular intervals of 10 minutes to collect all organic and promoted hashtags trending in the United States between January and April 2013, for a total of $N = 927$ hashtags. This constitutes our ground-truth dataset of *promoted* and *organic* trends.

We extracted a sample of organic trends observed during the first two weeks of March 2013 for our analysis. While Twitter allows for at most one promoted hashtag per day, dozens of organic trends appear in the same period. As a result, our dataset is highly imbalanced, with the promoted class more than ten times smaller than the the organic one (cf. Table 5.1). Such an imbalance, however, reflects our expectation to observe in the Twitter stream a minority of promoted conversations blended in a majority of organic content. Therefore we did not balance the classes by resampling, to study the campaign detection problem under realistic conditions.

Hashtags may trend multiple times on Twitter. However, those in our collection only trended once during our observation period. For each trend, we retrieved all tweets containing the trending hashtag from an archive containing a 10% random sample of the public Twitter stream. The collection period was hashtag-specific: for each hashtag we obtained all tweets produced in a four-day interval, starting two days before its trending point and extending to two days after that. This procedure provides an extensive coverage of the temporal history of each trending hashtag in our dataset and its related tweets, allowing us to study the characteristics of each trend before, during, and after the trending point.

Given that each trend is described by a collection of tweets over time, we can aggregate data in sliding time windows $[t, t + \ell)$ of duration ℓ and compute features on the subsets of tweets produced in these windows. A window can slide by time intervals of duration δ. The next window therefore contains tweets produced in the interval $[t + \delta, t + \ell + \delta)$. We experimented with various time window lengths and sliding parameters, and the optimal

performance is often obtained with windows of duration $\ell = 6$ hours sliding by $\delta = 20$ minutes.

We have made the IDs of all tweets involved in the trending hashtags analyzed in this paper available in a public dataset (`carl.cs.indiana.edu/data/ovarol/trend-dataset.tar.gz`).

5.2.2 Features

Our framework computes features from a collection of tweets in some time interval. The system generates 487 features in five different classes: network structure and information diffusion patterns, content and language, sentiment, timing, and user meta-data. The classes and types of features are reported in Table 5.2 and discussed next. All of the feature time series in this study are available in our public dataset.

5.2.2.1 Network and Diffusion Features

Twitter actively fosters interconnectivity. Users are linked by means of *follower/followee* relations. Content travels from person to person via *retweets*. Tweets themselves can be addressed to specific users via *mentions*. The network structure carries crucial information for the characterization of different types of communication. In fact, the usage of network features significantly helps in tasks like astroturf detection [240]. Our system reconstructs three types of networks: retweet, mention, and hashtag co-occurrence networks. Retweet and mention networks have users as nodes, with a directed link between a pair of users that follows the direction of information spreading — toward the user retweeting or being mentioned. Hashtag co-occurrence networks have undirected links between hashtag nodes when two hashtags have occurred together in a tweet. All networks are weighted according to the number of interactions and co-occurrences. For each network, a set of features is

Table 5.2: List of 487 features extracted by our framework.

Network (†)	Number of nodes	Number of edges
	(*) Strength distribution	(*) In-strength distribution
	(*) Out-strength distribution	(*) Distribution of connected components size
	Network density	Density of the largest connected component
	Mean shortest path length of the LCC	
User	(*) Sender's follower count	(*) Sender's followee count
	(*) Sender's number of favorite tweets	(*) Sender's number of statuses
	(*) Sender's number of lists subscribed to	(*) Originator's follower count
	(*) Originator's followee count	(*) Originator's number of favorite tweets
	(*) Originator's number of Twitter statuses	(*) Originator's number of lists subscribed to
Timing	Number of tweets appeared in a given window	(*) Time between two consecutive tweets
	(*) Time between two consecutive retweets	(*) Time between two consecutive mentions
Content	(*) Number of hashtags in a tweet	(*) Number of mentions in a tweet
	(*) Number of URLs in a tweet	(*,**) Frequency of POS tags in a tweet
	(*,**) Proportion of POS tags in a tweet	(*) Number of words in a tweet
	(*) Entropy of words in a tweet	
Sentiment	(***) Happiness scores of aggregated tweets	(***) Valence scores of aggregated tweets
	(***) Arousal scores of aggregated tweets	(***) Dominance scores of single tweets
	(*) Happiness score of single tweets	(*) Valence score of single tweets
	(*) Arousal score of single tweets	(*) Dominance score of single tweets
	(*) Polarization score of single tweets	(*) Entropy of polarization scores of single tweets
	(*) Pos. emoticons entropy of single tweets	(*) Neg. emoticons entropy of single tweets
	(*) Emoticons entropy of single tweets	(*) Ratio between pos. and neg. score of single tweets
	(*) Number of pos. emoticons in single tweets	(*) Number of neg. emoticons in single tweets
	(*) Total number of emoticons in single tweets	Ratio of tweets that contain emoticons

† We consider three types of network: retweet, mention, and hashtag co-occurrence networks.

* Distribution types. For each distribution, the following eight statistics are computed and used as individual features: min, max, median, mean, std. deviation, skewness, kurtosis, and entropy.

** Part-of-Speech (POS) tag. There are nine POS tags: verbs, nuns, adjectives, modal auxiliaries, pre-determiners, interjections, adverbs, wh-, and pronouns.

*** For each feature we compute mean and std. deviation.

computed, including in- and out-strength (weighted degree) distribution, density, shortest-path distribution, and so on. (cf. Table 5.2).

5.2.2.2 User-based Features

User meta-data is crucial to classify communication patterns in social media [115, 206]. We extract user-based features from the details provided by the Twitter API about the author of each tweet and the originator of each retweet. Such features include the distribution of follower and followee numbers, and the number of tweets produced by the users (cf. Table 5.2).

5.2.2.3 Timing Features

The temporal dimension associated with the production and consumption of content may reveal important information about campaigns and their evolution [129]. The most basic time-related feature we considered is the number of tweets produced in a given time interval. Other timing features describe the distributions of the intervals between two consecutive events, like two tweets or retweets (cf. Table 5.2).

5.2.2.4 Content and Language Features

Many recent papers have demonstrated the importance of content and language features in revealing the nature of social media conversations [47, 90, 184, 197, 209]. For example, deceiving messages generally exhibit informal language and short sentences [50]. Our system extracts language features by applying a *Part-of-Speech* (POS) tagging technique, which identifies different types of natural language components, or *POS tags*. The following POS tags are extracted: verbs, nouns, adjectives, modal auxiliaries, pre-determiners, interjections, adverbs, pronouns, and wh-pronouns (for details and examples see `www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html`). Tweets can be therefore analyzed to study how such POS

tags are distributed. Other content features include the length and entropy of the tweet content (cf. Table 5.2).

5.2.2.5 Sentiment Features

Sentiment analysis is a powerful tool to describe the attitude or mood of an online conversation. Sentiment extracted from social media conversations has been used to forecast offline events, including elections and financial market fluctuations [42, 272], and is known to affect information spreading [119, 207]. Our framework leverages several sentiment extraction techniques to generate various sentiment features, including *happiness score* [172], *arousal, valence and dominance scores* [295], *polarization and strength* [305], and *emotion score* [4] (cf. Table 5.2).

5.2.3 Feature Selection

Our system generates a set I of $|I| = 487$ features (cf. Table 5.2) designed to extract signals from a collection of tweets and distinguish promoted trends from organic ones. Some features are more predictive than others; some are by definition correlated with each other due to temporal dependencies. Most of the correlations are related to the volume of data. Analysis of feature correlations illustrated in Fig. 5.4. As we can see, many pairs of features are highly correlated. For instance the two most correlated features immediately prior to the trending point are the size of the hashtag cooccurrence network and the size of its largest connected component (Pearson's $\rho = 0.75$). This is why it is important to perform feature selection to eliminate redundant features and identify a combination of features that yield good classification performance.

There are several methods to select the most predictive features in a classification task [146]. We implemented a simple greedy forward feature selection method, summarized as

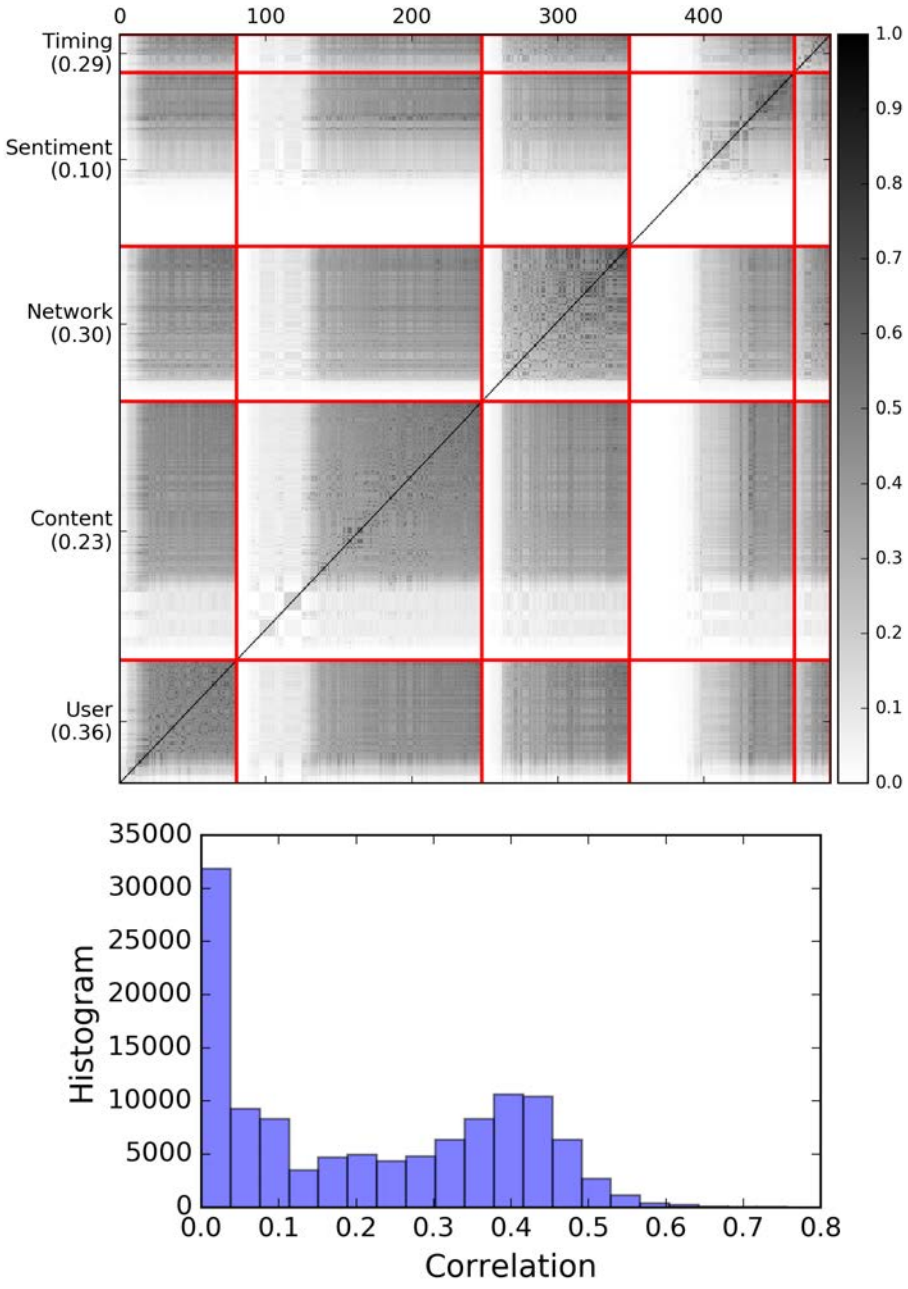

Figure 5.4: Pairwise correlation between features averaged across trends (top) and histogram of correlation values (bottom).

follows: (i) initialize the set of selected features $S = \emptyset$; (ii) for each feature $i \in I - S$, consider the union set $U = S \cup \{i\}$; (iii) train the classifier using the features in U; (iv) test the average performance of the classifier trained on this set; (v) add to S the feature that provides the best performance; (vi) repeat (ii)–(v). We terminate the feature selection procedure if the AUC (cf. Sec. 5.2.5) increases by less than 0.05 between two consecutive steps. Most of the experiments terminate after selecting fewer than 10 features. The time series for the selected features are passed as input to the learning algorithms. In the next subsections we provide details about our experimental setting and learning models.

5.2.4 Experimental Setting

Our experimental setting follows a pipeline of feature selection, model building, and performance evaluation. We apply the *wrapper* approach to select features and evaluate performance iteratively [161]. During each iteration (Fig. 5.5), we train and evaluate models using candidate subsets of features and expand the set of selected features using the greedy approach described in Sec. 5.2.3. Once we identify the set of features that performs best, we report results of experiments using only this set of features.

In each experiment and for each feature, an algorithm receives in input a time series with $L = 35$ data points to carry out its detection. The length of the time series and its delay D with respect to the trending point are discussed in Sec. 5.3; different experiments will consider different delays.

A set of feature time series is used to either train a learning model or evaluate its accuracy. The learning algorithms are discussed in the next subsection. For evaluation, we compute a Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) versus the false positive rate (FPR) at various thresholds. Accuracy is evaluated by measuring the Area Under the ROC Curve (AUC) [112] with 10-fold cross validation,

and averaging AUC scores across the folds. A random-guess classifier produces the diagonal line where TPR equals FPR, corresponding to a 50% AUC score. Classifiers with higher AUC scores perform better and the perfect classifier in this setting achieves a 100% AUC score. We adopt AUC to measure accuracy because it is not biased by the imbalance in our classes (75 promoted trends versus 852 organic ones, as discussed earlier).

5.2.5 Learning Algorithms

Let us describe the learning systems for online campaign detection based on multidimensional time-series data from social media. We identified an algorithm, called *K-Nearest Neighbor with Dynamic Time Warping* (KNN-DTW), that is capable of dealing with multidimensional time series classification. For evaluation purposes, we compare the classification results against two baselines: SAX-VSM and KNN. These three methods are described next.

5.2.5.1 KNN-DTW Classifier

KNN-DTW is a state-of-the-art algorithm to classify multidimensional time series, illustrated in Fig. 5.5. During learning, we provide our model with training and testing sets generated by 10-fold cross validation. Time series for each feature are processed in parallel using *dynamic time warping* (DTW), which measures the similarity between two time series after finding an optimal match between them by "warping" the time axis [33]. This allows the method to absorb some non-linear variations in the time series, for example different speed or resolution of the data.

For efficiency, we initially apply a time series coarsening strategy called *piece-wise aggregation*. We split each original time series into p equally long sections and replace the time-series values by the section averages, reducing the dimensionality from L to $L' = L/p$. For trend i and feature k, we thus obtain a coarsened time series $f_k^i = \{f_{k,1}^i, f_{k,2}^i, \cdots, f_{k,L'}^i\}$.

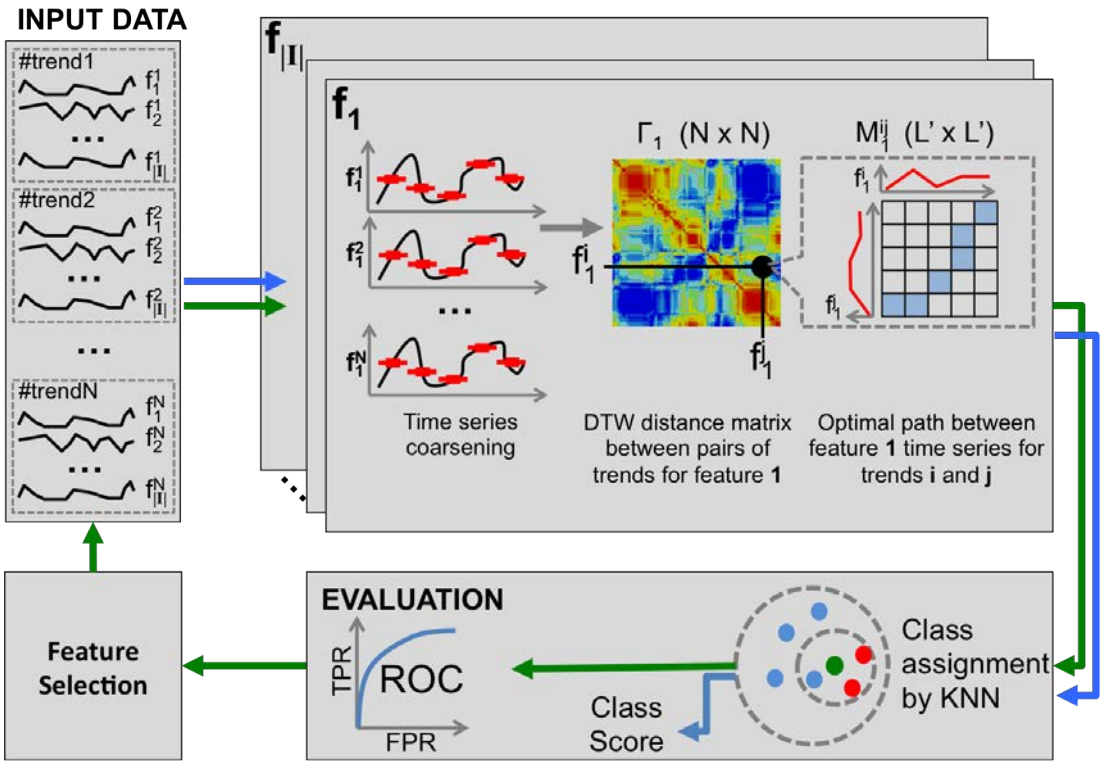

Figure 5.5: Wrapper method description for KNN-DTW. We present the pipeline of our complete system, including feature selection and model evaluation steps. Input data feed into the system for training (green arrow) and testing (blue arrow) steps.

Then, DTW computes the distance between all pairs of points of two given trend time series f_k^i and f_k^j. Each element of the resulting $L' \times L'$ distance matrix is $M_k^{ij}(t,t') = (f_{k,t}^i - f_{k,t'}^j)^2$. Points closer to each other are more likely to be matched. To create a mapping between the two time series, an optimal path is computed over the time-series distance matrix. A path must start from the beginning of each time series and terminate at its end. The path between first and last points is then computed by minimizing the cumulative distance (γ) over alternative paths. This problem can be solved via dynamic programming [33] using the following recurrence: $\gamma(t,t') = M(t,t') + min\{\gamma(t-1,t'-1), \gamma(t-1,t'), \gamma(t,t'-1)\}$ (indices i,j,k dropped for readability). The distance γ_k^{ij} is used as the ij-th element of the $N \times N$ trend similarity matrix Γ_k.

The computation of similarity between time series using DTW requires $O(L'^2)$ operations. Some heuristic strategies use lower-bounding techniques to reduce the computational complexity [169]. Another technique is to re-sample the data before adopting DTW. Our coarsening approach reduces the computational costs by a factor of p^2. We achieved a significant increase in efficiency with marginal classification accuracy deterioration by setting $p = 5$ ($L' = 7$).

In the evaluation step, we use the K-Nearest Neighbor (KNN) algorithm [86] to assign a class score to a test trend q. We compare q with each training trend i to obtain a DTW distance γ_k^{iq} for each feature k. We then find the $K = 5$ labeled trends with smallest DTW distance from q, and compute the fraction of promoted trends s_k^q among these nearest neighbors. We finally average across features to obtain the class score $\bar{s^q}$. Higher values of $\bar{s^q}$ indicate a high probability that q is a promoted trend. Class scores, together with ground-truth labels, allow us to compute the AUC of a model, which is then averaged across folds according to cross validation.

5.2.5.2 SAX-VSM Classifier

Our first baseline, called SAX-VSM, blends symbolic dimensionality reduction and vector space models [251]. Time series are encoded via Symbolic Aggregate approXimation (SAX), yielding a compact symbolic representation that has been used for time series anomaly and motif detection, time series clustering, indexing, and more [188, 189]. A symbolic representation encodes numerical features as words. A vector space model is then applied to treat time series as documents for classification purposes, similarly to what is done in information retrieval. In our implementation, we first apply piece-wise aggregation and then use SAX to represent the data points in input as a single word of L' letters from an alphabet $\aleph$. This choice and the parameters $|\aleph| = 5$ and $L' = 4$ are based on prior optimization [251], and variations to these settings only marginally affect performance. Each time-series value is mapped into a letter by dividing the range of the feature values into $|\aleph|$ regions in such a way as to obtain equiprobable intervals under an assumption of normality [189]. In the training phase, for each feature, we build two sets of words corresponding to organic and promoted trends, respectively. In the test phase, a new instance is assigned to the class with the majority of word matches across features. In case of a tie we assign a random class. For further details about this baseline and its implementation, we refer the reader to the SAX-VSM project website (`github.com/jMotif/sax-vsm_classic`).

5.2.5.3 K-Nearest Neighbors Classifier

Our second baseline is an off-the-shelf implementation of the traditional *K-Nearest Neighbors* algorithm [86] for time-series classification. We used the Python scikit-learn package [232]. We selected KNN because it can capture and learn time-series patterns without requiring any pre-processing of the raw time-series data. We created the feature vectors for each trend by concatenating into a single vector the continuous-valued time series representing

each feature. The nearest neighbor classifier computes the Euclidean distance between pairs of single-vector time series. For a test trend, the class score is given by the fraction of promoted trends among the $K = 5$ nearest neighbors.

5.3 Results

In this section, we present results of experiments design to evaluate the ability of our machine learning framework to discriminate between organic and promoted trends. For all experiments, each feature time series consists of 120 real-valued data points equally divided before and after the trending point. Although in principle we could use the entire time series for classification, ex-post information would not serve our goal of early detection of social media campaigns in a streaming scenario that resembles a real setting, where information about the future evolution of a trend is obviously unavailable. For this reason, we consider only a subset consisting of L data points ending with delay D since the trending point; $D \leq 0$ data points for early detection, $D > 0$ for classification after trending. We evaluate the performance of our detection framework as a function of the delay parameter D. The case $D = 0$ involves detection immediately at trending time. However, we also consider $D < 0$ to examine the performance of our algorithms based on data preceding the trending point; of course the detection would not occur until $D = 0$, when one would become aware of the trending hashtag. Time series are encoded using the settings described above ($L = 35$ windows of length $\ell = 6$ hours sliding every $\delta = 20$ minutes).

5.3.1 Method Comparison

We carried out an extensive benchmark of several configurations of our system for campaign detection. The performance of the algorithms as a function of varying delays D is plotted in Fig. 5.6.

In addition, we introduce random temporal shifts for each trend time series to test the robustness of the algorithms. In real-world scenarios, one would ideally expect to detect a promoted trend without knowing the trending point. To simulate such scenarios, we designed an experiment that introduces variations that randomly shift each time series around its trending point. The temporal shifts are sampled from gaussian distributions with different variances. We present the results of this experiment in Fig. 5.7.

KNN-DTW and KNN display the best detection accuracy (measured by AUC) in general. Their performance is comparable (Fig. 5.6). The AUC score is on average around 95% for detecting promoted trends after trending. In the early detection task, we observe scores above 70%. This is quite remarkable given the small amount of data available before the trending point. KNN-DTW also displays a strong robustness to temporal shifts, pointing to the advantage of time warping (Fig. 5.7). The KNN algorithm is less robust because it computes point-wise similarities between time series without any temporal alignment; as the variance of the temporal shifts increases, we observe a significant drop in accuracy. SAX-VSM benefits from the time series encoding and provides good detection performance (on average around 80% AUC) but early detection accuracy is poor, close to random for $D < 0$. A strong feature of SAX-VSM is its robustness to temporal shifts, similar to KNN-DTW.

Our experiments suggest that temporal encoding is a crucial ingredient for successful classification of time-series data. Encoding reduces the dimensionality of the signal. More importantly, encoding preserves (most) information about temporal trends and makes an algorithm robust to random shifts, which is an importance advantage in real-world scenarios. SAX-VSM ignores long-term temporal ordering. KNN-DTW, on the other hand, computes similarities using a time series representation that preserves the long-term temporal order, even as time warping may alter short-term trends. This turns out to be a crucial advantage to achieve both high accuracy and robustness.

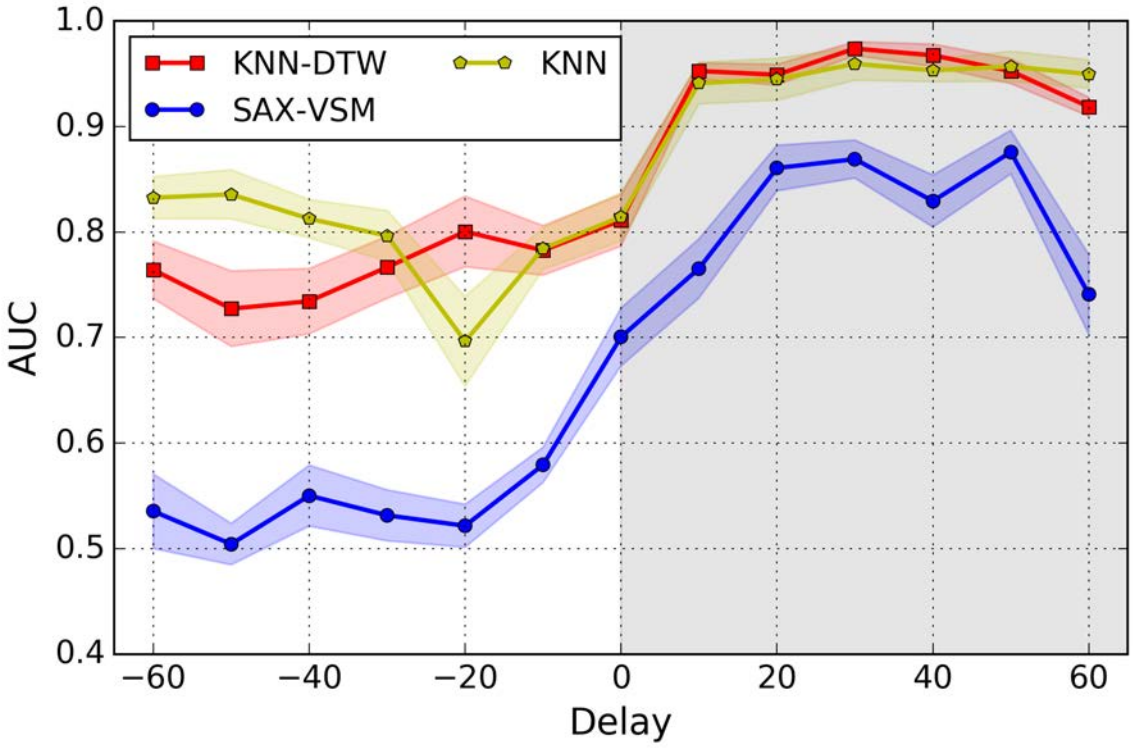

Figure 5.6: Methods comparison. Classification performance of different learning algorithms on encoded and raw time series. The AUC is measured for various delays D. Confidence intervals represent standard errors based on 10-fold cross validation.

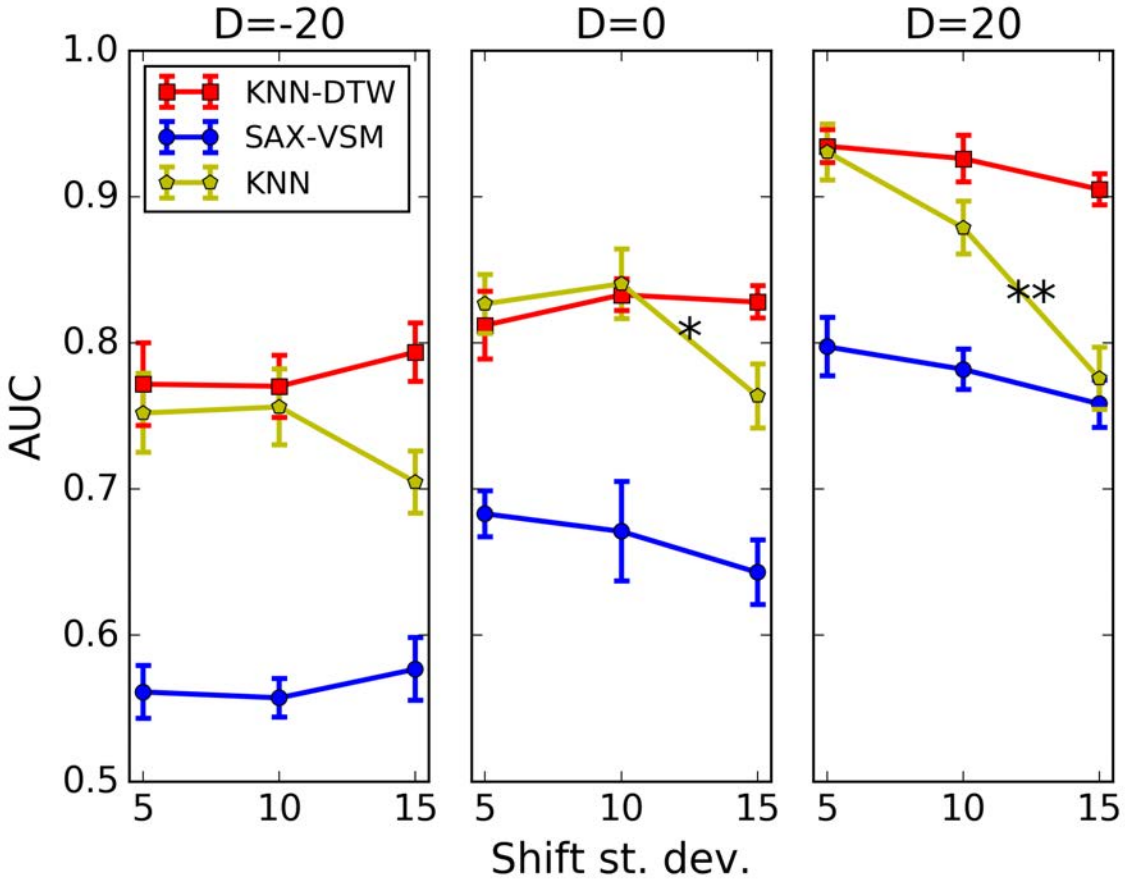

Figure 5.7: Temporal robustness. AUC of different learning algorithms with random temporal shifts versus the standard deviation of the shifts. We repeated the experiment for various delay values D. Significance levels of differences in consecutive experiments are marked as (*) $p < 0.05$ and (**) $p < 0.01$.

Using AUC as an evaluation metric has the advantage of not requiring discretization of scores into binary class labels. However, detection of promoted trends in real scenarios requires binary classification by a threshold. In this way we can measure accuracy, precision, recall, and identify misclassified accounts. Fig. 5.8 illustrates the distribution of probabilistic scores produced by the KNN-DTW classifier as a function of the delay for the two classes of trends, organic and promoted. The scores are computed for leave-out test instances, across folds. An ideal classifier would separate these distributions completely, achieving perfect accuracy. Test instances in the intersection between two distributions either are misclassified or have low-confidence scores. Examples of misclassified instances are discussed in Sec. 5.3.3. For $D < 0$, KNN-DTW generates more conservative scores, and the separation between the organic and promoted class distributions is smaller. For $D > 0$, KNN-DTW scores separate the two classes well. To convert continuos scores into binary labels, we calculated the threshold values that maximize the F1 score of each experiment; this score combines precision and recall. Trends with scores above the threshold are labeled as promoted. The best accuracy and F1 score are obtained shortly after trending, at $D = 20$.

5.3.2 Feature Analysis

Let us explore the roles and importance of different features for trend detection. To this end, we identify the significant features using the greedy selection algorithm described in Sec. 5.2.3, and group them by the five classes (user meta-data, content, network, sentiment, and timing) previously defined. We focus on KNN-DTW, our best performing method. After selecting the top 10 features for different delays D, we compute the fractions of top features in each class, as illustrated by Fig. 5.9. We list the top features for experiments $D = 0$ (early detection) and $D = 40$ (classification) in Table 5.3.

The usefulness of content features does not appear to change significantly between early

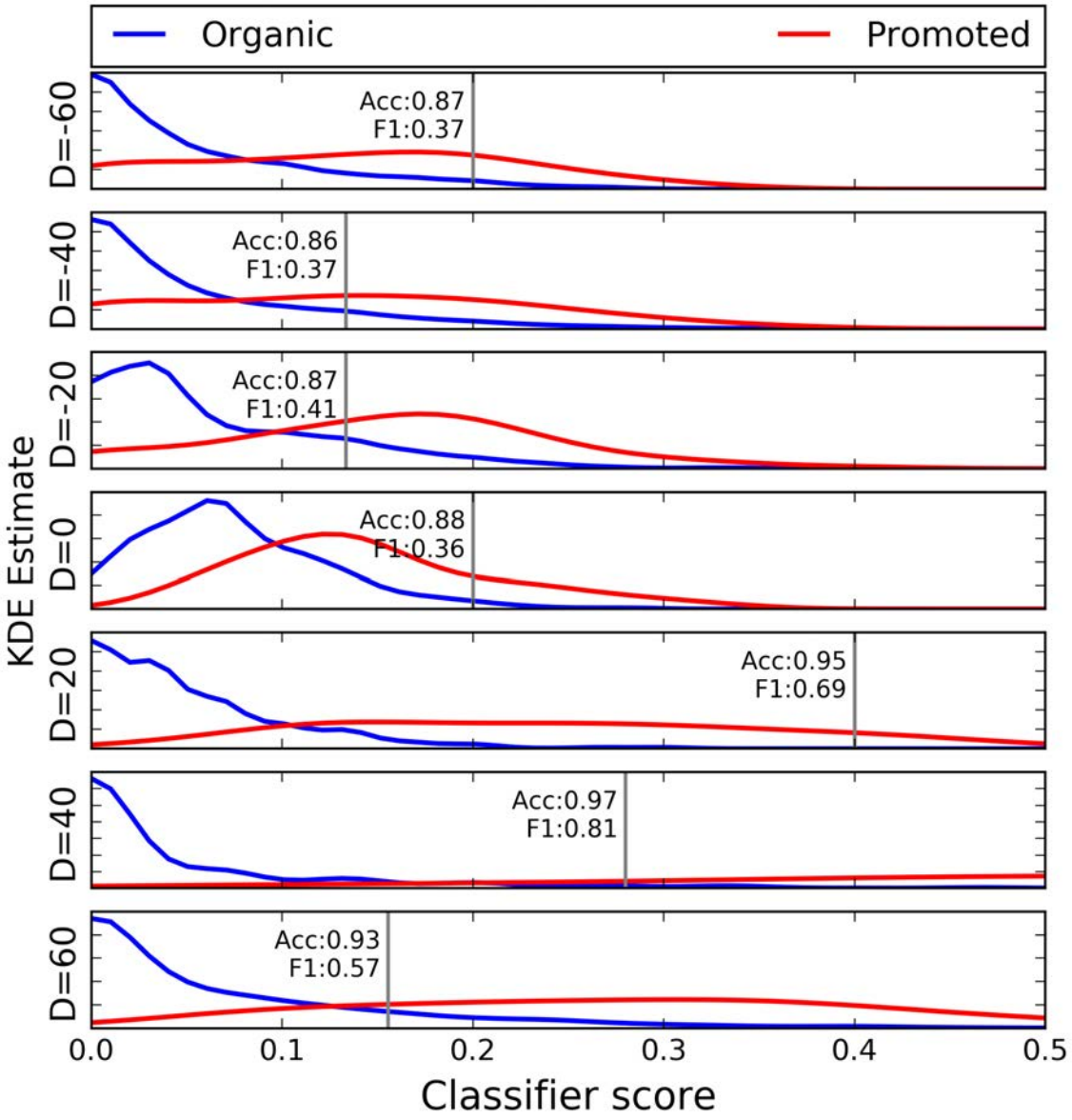

Figure 5.8: Distributions of KNN-DTW classifier scores. We use Kernel Density Estimation (KDE), a non-parametric smoothing method, to estimate the probability densities based on finite data samples. We also show the threshold values that separate the two classes yielding an optimal F1 score.

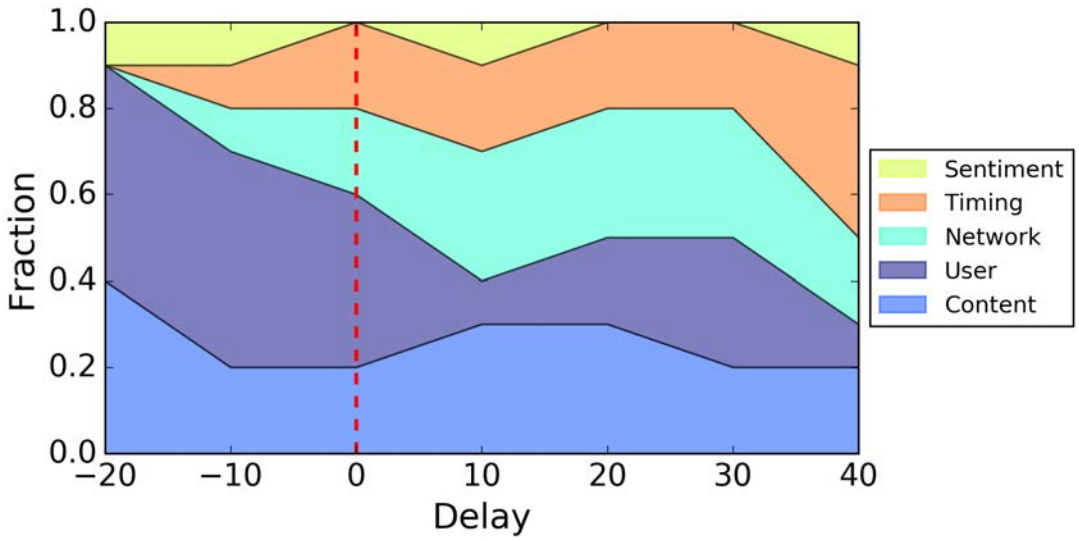

Figure 5.9: KNN-DTW feature analysis. Stacked plot showing how different feature classes are represented among the top 10 selected features.

and late detection. In the early detection task, user features seem to contribute significantly more than any other class, possibly because early adopters reveal strong signals about the nature of trends. As we move past the trending point, signals from early adopters are flooded by increasing numbers of participants. Timing and network features become increasingly important as the involvement of more users allows to analyze group activity and network structure patterns.

5.3.3 Analysis of Misclassifications

We conclude our analysis by discussing when our system fails. In Fig. 5.10, we illustrate how some key features of misclassified trends diverge from the majority of the trends that are correctly classified. We observe that some misclassified trends follow the temporal characteristics of the other class. This is best illustrated in the case of volume (number of tweets).

An advantage of continuous class scores is that we can tune the classification threshold to achieve a desired balance between precision and recall, or between false positives and

Table 5.3: Top 10 features for experiments with different values of D.

Delay	Features	Classes
40	Number of tweets	Timing
	Max. proportion of pronouns in a tweet	Content
	Entropy of hashtag cooccurrence network degree	Network
	Entropy of time between two consecutive mentions	Timing
	Mean time between two consecutive tweets	Timing
	Entropy of emoticon scores	Sentiment
	Median time between two consecutive tweets	Timing
	Max. originator's followers count	User
	Kurtosis of mention network degree distribution	Network
	Entropy of pre-determiner POS frequency in a tweet	Content
0	Max. hashtag cooccurrence network degree	Network
	Entropy of number of originator's friends count	User
	Max. originator's statuses count	User
	Median time between two consecutive tweets	Timing
	Skewness of time between two consecutive mentions	Timing
	Median of sender's lists count	User
	Min. originator's lists count	User
	Median of mention network out-degree	Network
	Min. frequency of adjective POS in a tweet	Content
	Mean frequency of noun POS in a tweet	Content

false negatives. False negative errors are the most costly for a detection system: a promoted trend mistakenly labeled as organic would easily go unchecked among the larger number of correctly labeled organic trends. Focusing our attention on a few specific instances of false negatives generated by our system, we gained some insight on the reasons triggering the mistakes. First of all, it is conceivable that promoted trends are sustained by organic activity before promotion and therefore they are essentially indistinguishable from organic ones until the promotion triggers the trending behavior. It is also reasonable to expect a decline in performance for long delays: as more users join the conversation, promoted trends become harder to distinguish from organic ones. This may explain the dip in accuracy observed for the longest delay (cf. Fig. 5.6).

False positives (organic trends mistakenly labeled as promoted) can be manually filtered out in post-processing and are therefore less costly. However, analysis of false positives provides for some insight as well. Some trends in our dataset, such as `#watchsuitstonight` and `#madmen`, were promoted via alternative communication channels (television and radio), rather than via Twitter. This has become a common practice in recent years, as more and

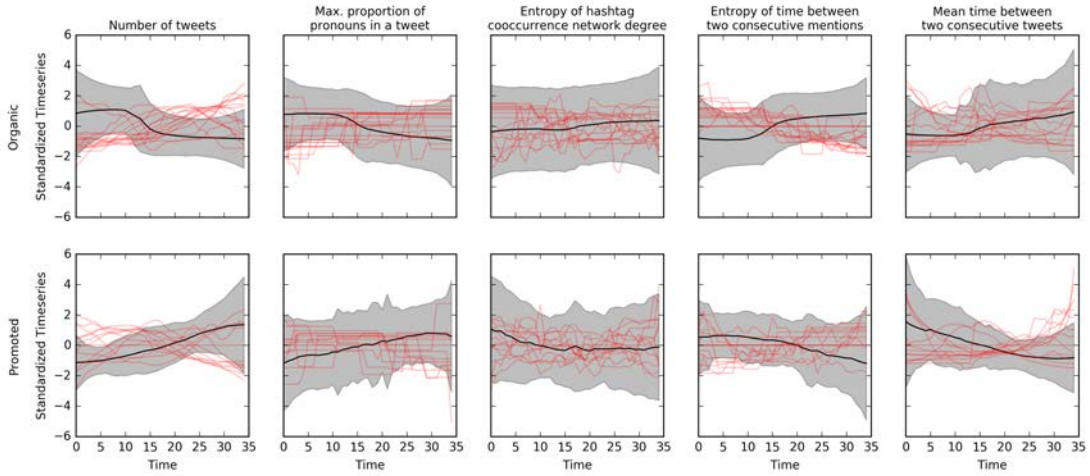

Figure 5.10: Comparison between feature time series of misclassified and correctly classified trends. Time series of the top five features (columns) for promoted (top) and organic (bottom) trends in the $D = 40$ detection task. The black lines and gray areas represent the average and 95% confidence intervals of time series for correctly classified trends. Time series of misclassified trends are shown in red. Misclassified organic trends (false positives) are: `#whyiwatchsuits`, `#watchsuitstonight`, `#bobsantigoldlive`, `#evildead`, `#galaxyfamily`, `#gethappy`, `#madmen`, `#makeboringbrilliant`, `#nyias`, `#oneboston`, `#stingray`, `#thewalkingdead`, and `#timeto365`. Misclassified promoted trends (false negatives) are: `#1dmemories`, `#8thseed`, `#20singersthatilike`, `#mentionsomeonecuteandbeautiful`, `#bnppo13`, `#ciaa`, `#expowest`, `#jaibrooksforpresident`, `#justintimberweek`, `#kobalt400`, `#nyc`, `#realestate`, `#stars`, `#sxsw`, `#wbc`, and `#wcw`.

more Twitter campaigns are mentioned or advertised externally to trigger organic-looking responses in the audience. Our system recognized such instances as promoted, whereas their ground-truth labels did not. Those campaigns were therefore wrongly counted as false positives, penalizing our algorithms in the evaluation. We find it remarkable that in these cases our system is capable of learning the signature of promoted trends, even though the promotion occurs outside of the social media itself.

5.4 Related Work

Recent work on social media provides a better understanding of human communication dynamics such as collective attention and information diffusion [296], the emergence of

trends [116, 183], social influence and political mobilization [44, 83, 84, 284].

Different information diffusion mechanisms may determine the trending dynamics of hashtags and other memes on social media. Exogenous and endogenous dynamics produce memes with distinctive characteristics [116, 119, 181, 216, 263]: external events occurring in the real world (e.g., a natural disaster or a terrorist attack) can generate chatter on the platform and therefore trigger the trending of a new, unforeseen hashtag; other topics (e.g., politics or entertainment) are continuously discussed and sometimes a particular conversation can accrue lots of attention and generate trending memes. The promotional campaigns studied here can be seen as a type of exogenous factor affecting the visibility of memes.

The present work, to the best of our knowledge, is the first to investigate the early detection of promoted content on social media. We focus our attention on advertisement, which can play an important role in information campaigns. Trending memes are considered an indicator of collective attention in social media [181,308], and as such they have been used to predict real-world events, like the winner of a popular reality TV show [75]. Although emerging from collective attention, communication on social media can be manipulated, for example for political gain, as in the case of astroturf [204, 240].

Recent work analyzes emerging topics, memes, and conversations triggered by real world events [5, 23, 60]. Studies of information dissemination reveal mechanisms governing content production and consumption [73] as well as prediction of future content popularity. Cheng *et al.* study the prediction of photo-sharing cascade size [65] and recurrence [66] on Facebook. Machine learning models can predict future popularity of emerging hashtags and content on social media [191, 270]. Features extracted from content [157], sentiment [119, 173], community structure [298, 299], and temporal signatures [120, 234, 294] are commonly used to train such models. In this paper we leverage similar features, but for the novel task of campaign detection. Furthermore, our task is more challenging because we deal with

dynamic features whose changes over time are captured in high-dimensional time series.

Another topic related to our research is rumor detection. Rumors may emerge organically as genuine conversation and spread out of control. They are characterized and sustained by ambiguous contexts, where correctness and completeness of information or the meaning of a situation is not obviously apparent [103]. Examples are situations of crisis or topics of public debate [202]. Existing systems to identify rumors are based mostly on content analysis [176, 235] and clustering techniques [114, 156]. An open question is to determine if rumor detection might benefit from the wide set of feature classes we propose here.

The proposed framework is based on a mixture of features common in social media data, including emotional and sentiment information. The literature has reported extensively on the use of social media content to describe emotional and demographic characteristics of users [119, 206, 207]. The use of language in online communities is the focus of two recent papers [90, 197]: the authors observe that the language of social media users evolves, and common patterns emerge over time. The language style of users adapts to achieve better fitness in the conversation [92]. These findings suggest that language contains strong signals, in particular if studied in conjunction with other dimensions of the data. Our study confirms the importance of content for campaign detection.

Finally, our system builds on network features and diffusion patterns of social media messages. Network structure and information diffusion in social media have been studied extensively [15, 182]. Network features are highly predictive of certain types of social media abuse, like astroturf, that attempt to simulate grassroots online conversations [115, 240, 241, 282]. Such artificial campaigns produce peculiar patterns of information diffusion: the topology of retweet or mention networks is often a stronger signal than content or language. The present findings are consistent with this body of work, as network features are helpful in detecting promoted content after trending.

5.5 Conclusions

As we increasingly rely on social media to satisfy our information needs, it is important to recognize the dynamics behind online campaigns. In this paper, we posed the problem of early-detection of promoted trends on social media, discussed the challenges that this problem presents, and proposed a supervised computational framework to attack it. The proposed system leverages time series representing the evolution of different features characterizing trending campaigns. The list includes features relative to network structure and diffusion patterns, sentiment, language and content features, timing, and user meta-data. We demonstrated the crucial advantages of encoding temporal sequences.

We achieved good accuracy in campaign detection. Our early detection performance is remarkable when one considers the challenging nature of the problem and the low volume of data available in the early stage of a campaign. We also studied the robustness of the proposed algorithms by introducing random temporal shifts around the trending point, simulating realistic scenarios in which the trending point can only be estimated with limited accuracy.

One of the advantages of our framework is that of providing interpretable feature classes. We explored how content, network, and user features affect detection performance. Extensive feature analysis revealed that signatures of campaigns can be detected early, especially by leveraging content and user features. After the trending point, network and temporal features become more useful.

The availability of data about organic and promoted trends is subject to Twitter's recipe for selecting trending hashtags. There is no certain way to know if and when social media platforms make any changes to such recipes. However, nothing in our approach assumes any knowledge of a particular platform's trending recipe. If the recipe changes, our system could be retrained accordingly.

This work represents an important step toward the automatic detection of campaigns. The problem is of paramount importance, since social media shape the opinions of millions of users in everyday life. Further work is needed to study whether different classes of campaigns (say, legitimate advertising vs. terrorist propaganda) may exhibit characteristics captured by distinct features. Many of the features leveraged in our model, such as those related to network structure and temporal attributes, capture activity patterns that could provide useful signals to detect astroturf [240]. Therefore, our framework could in principle be applied to astroturf detection, if longitudinal training data about astroturf campaigns were available.

CHAPTER 6

Social Bots

Increasing evidence suggests that a growing amount of social media content is generated by autonomous entities known as social bots [115]. As opposed to social media accounts controlled by humans, bots are controlled by software, algorithmically generating content and establishing interactions. While not all social bots are harmful, there is a growing record of malicious applications of social bots. Some emulate human behavior to manufacture fake grassroots political support [240], promote terrorist propaganda and recruitment [31, 118], manipulate the stock market [113], and disseminate rumors and conspiracy theories [34].

Discussion of social bot activity, the broader implications on the social network, and the detection of these accounts are becoming central research avenues [46, 113, 115, 180]. The magnitude of the problem is underscored by a social bot detection challenge recently organized by DARPA to study information dissemination mediated by automated accounts and to detect malicious activities carried out by these bots.

In this section, I will describe: (i) our approach, ranked third worldwide, on detecting social bots for DARPA challenge [266]; (ii) BotOrNot system on social bot detection and estimation of social bot population on Twitter [94, 282]; (iii) analysis of Twitter human-bot ecosystem [282].

6.1 DARPA Social Bot Detection Challenge

In 2015, DARPA conducted a competition on social bot detection. Task of the challenge is to identify influence bots supporting pro-vaccination discussion on Twitter. During the task organic activity of the anti-vaccine community and automated posts by pro-vaccine bots are mixed.

In this challenge, our team designed a system to track, store and process the streaming data in real-time, while creating and updating the profiles of the accounts involved in the conversation, along with their corresponding features. As a result our team successfully identified all bots a week before the competition ended. We ranked as second fastest and third most accurate team worldwide [266].

Our approach consist of three steps:

- Extraction of user-based features
- Filtering search space based on various heuristics
- Human assisted inspection of suspicious users and activities through visualizations and interactive data exploration.

6.1.1 Feature Extraction

Our system builds a dynamic profile for each user participating in the conversation, for rapid data access, analysis, and classification. The system also generates feature vectors describing user profiles, updated every 6 hours, for classification purposes.

We adopt subset of features designed for our Twitter bot detection system *BotOrNot*. The features can be summarized in five classes: user metadata, content, sentiment, network, and temporal features. These features were carefully selected to reflect hand-crafted rules designed to identify suspicious activity. Examples of such rules include: (i) low entropy of topics of interest of the account, to identify thematically-focused users; (ii) anomalous levels

of retweets or mentions, to capture users attempting to attract attention; (iii) anomalous connectivity patterns, to detect suspicious cliques; (iv) coordinated attempts to address specific human users, to identify orchestrated targeting; (v) suspicious growth-rate in followers, following, or content production levels; (vi) suspicious temporal patterns, as opposed of natural human circadian activity; (vii) high-volume of near-duplicate content; (viii) high-degree of sentiment polarization; and (ix) interactions focused on users in the target population, as opposed to external users.

As the stream of data was “replayed", our system periodically re-computed the user feature vectors. The pairwise cosine similarity between the feature vectors highlights the most similar pairs of users. Once we started to identify bots in the conversation, matching the users most similar to the detected bots allowed for timely detection of new bots. In Fig. 6.1 we show the distribution of the pairwise cosine similarity between pairs of feature vectors characterizing bots, as opposed to bot-human pairs. The similarity between bots tends to be higher than between bots and humans. The bot-bot similarity exhibits a bimodal distribution that reflects the presence of two types of bots designed by two red teams: bots designed by same team are more similar to each other.

6.1.2 Heuristic Filtering

In the earlier stage of the competition, we developed various heuristic techniques to narrow the search space. Specifically, three strategies worked well: (i) analysis of the hashtag co-occurrence network; (ii) duplicate-image search; and (iii) dynamic tracking of network growth.

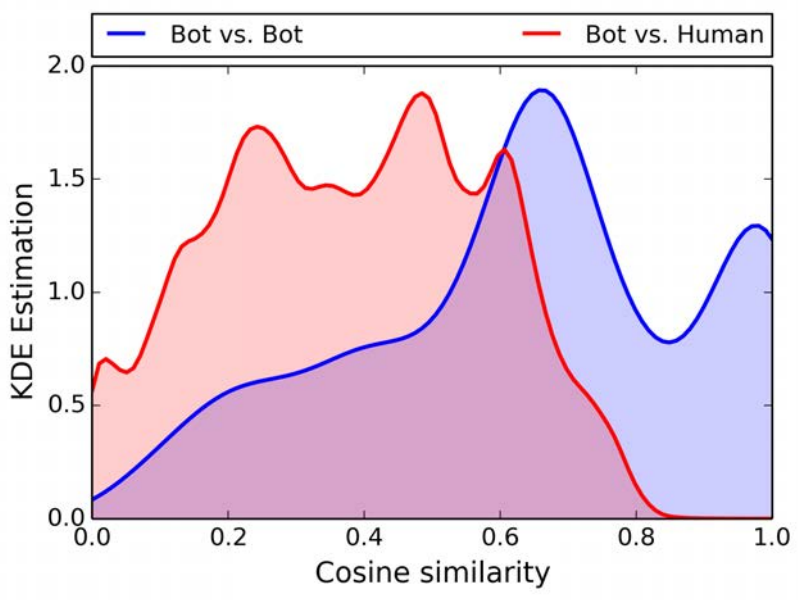

Figure 6.1: Distribution of cosine similarity between pairs of accounts.

6.1.2.1 Hashtag Co-occurrence Network

Starting from a provided list of vaccine-related hashtags, we collected all tweets appearing in the competition stream that contained at least one of those hashtags. The system constructed a hashtag co-occurrence network, where each node represents a unique hashtag and edges between two nodes are weighted by the number of times these two hashtag are observed together in a tweet (see Fig. 6.2).

Using the hashtag co-occurrence networks, we were able to identify other campaign-related hashtags to enrich the list of competition-relevant keywords. These were later used to separate users into categories of pro- and anti-vaccine. The proportion of tweets users posted containing any of these hashtags resulted in a strongly predictive feature.

6.1.2.2 Image Search

A common approach to create realistic bot profiles is to impersonate other users by cloning information such as descriptions, names, and profile pictures. We built an algorithm to

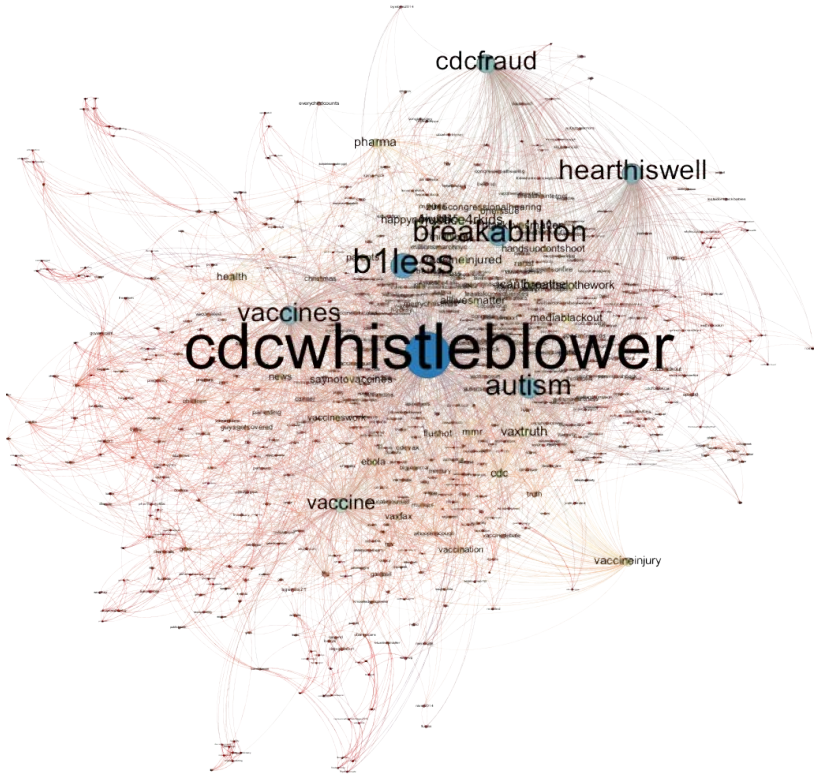

Figure 6.2: Hashtag co-occurrence network of vaccine discussions

detect duplicate user pictures using an online image search service[1]. Seven out of 39 bots were detected using this heuristics. These bots used images from the Wikipedia domain as their profile pictures.

6.1.2.3 Network Growth

In the competition dataset, a friendship network snapshot was provided every week. We studied the topological changes of these temporal networks and identified users that created suspicious levels of connections with anti-vaccine activists.

[1]www.tineye.com

6.1.3 Interactive Data Exploration

Information visualization is a crucial part of the our decision system. Expert knowledge is still required to conclude that a particular user is a social bot while limiting the number of false positives. We developed a web application similar to the Twitter platform to create and populate user profile information and timelines in real time (see Fig. 6.3). This interface includes charts to monitor temporal changes in user metadata, such as the number of followers, friends, and posts.

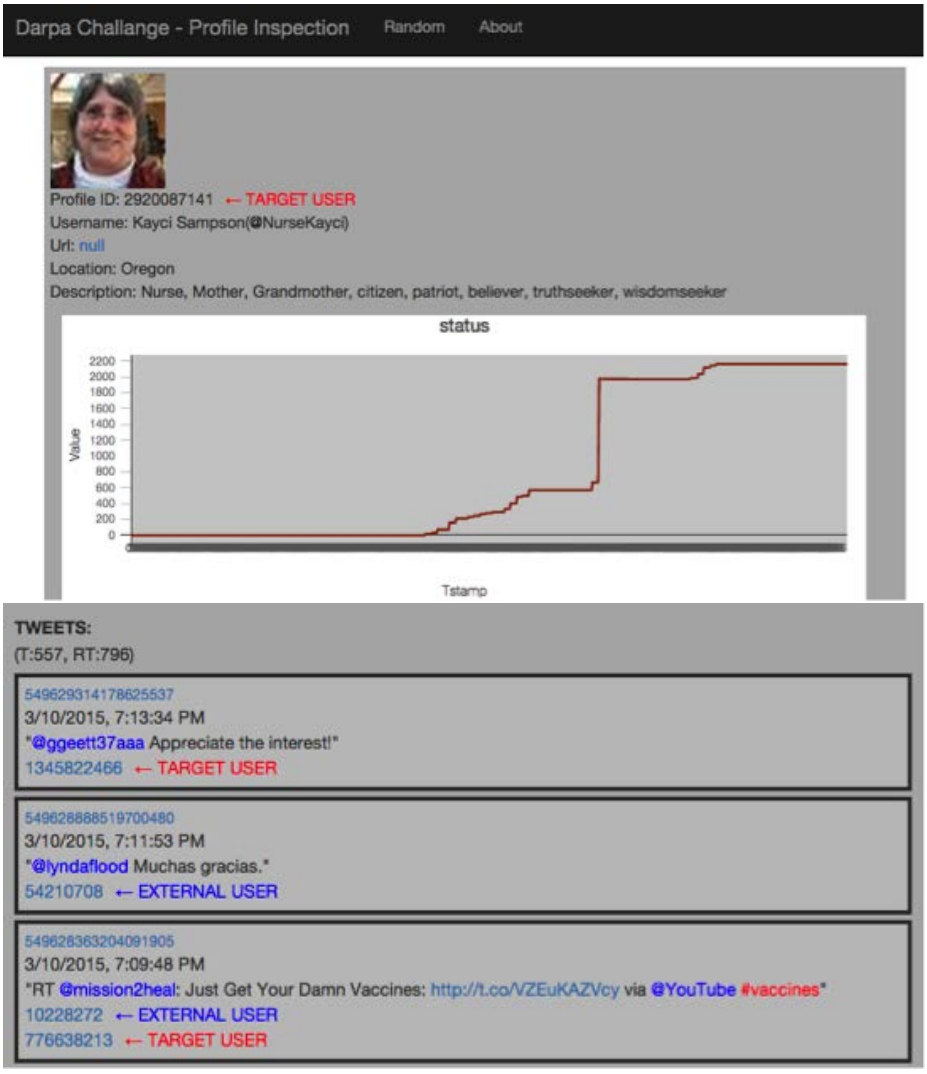

Figure 6.3: Interactive web interfaces designed to analyze user and content data.

6.2 BotOrNot: Social Bot Detection System

In this section, we present *BotOrNot*, our platform to evaluate whether a Twitter account is controlled by human or machine. This service is publicly available via the website[2] or via Python or REST APIs.[3,4] *BotOrNot* takes a Twitter screen name, retrieves that account's recent activity, then computes and returns a bot-likelihood score. For website users, this score is accompanied by plots of the various features used for prediction purposes as shown in Fig. 6.4.

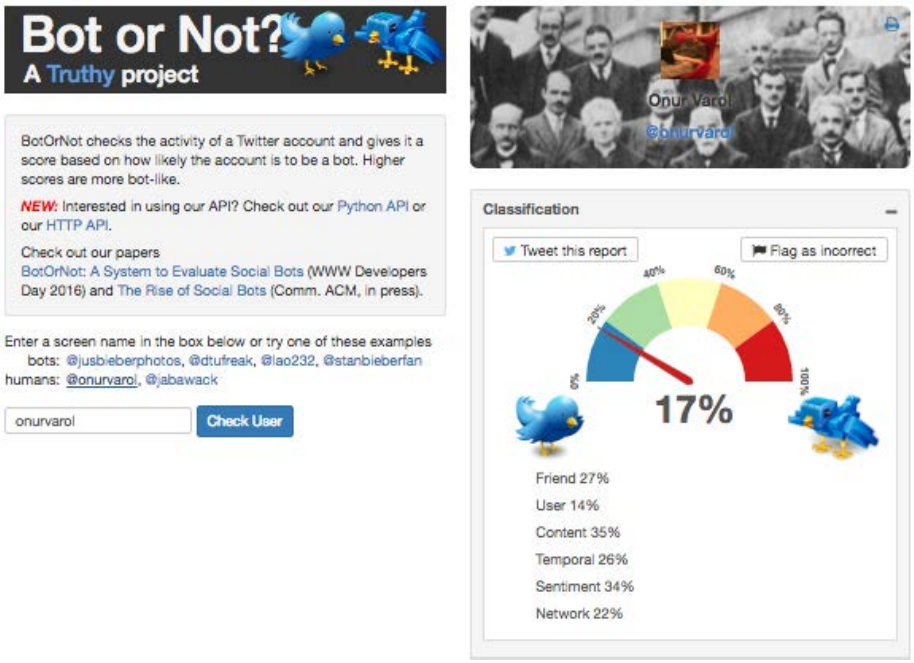

Figure 6.4: BotOrNot web interface

In the following part I would like to describe technical aspects of the system. We later used BotOrNot system to analyze Twitter ecosystem and estimate amount of social bots in the active Twitter population.

Let us first introduce our bot detection framework, which evaluates more than one thousand features from a target user's Twitter account, as well as from accounts of the target's

[2] `truthy.indiana.edu/botornot`
[3] `github.com/truthy/botornot-python`
[4] `truthy.indiana.edu/botornot/rest-api.html`

followers and followees (friends). These features are used to compare the target's behavior to that of known social bots. We later train and evaluate our initial model by using an available social bot dataset and off-the-shelf learning algorithms.

6.2.1 Feature Extraction

We distill 1,150 features in six different classes using the Twitter API. The classes and types of features are reported in Table 6.1 and described more in detail below.

6.2.1.1 User-based Features

Features extracted from user meta-data have been used to classify users and patterns before [115, 206]. We extract user-based features from meta-data available through the Twitter API. Such features include the number of friends and followers, the number of tweets produced by the users, profile description and settings (cf. Table 6.1).

6.2.1.2 Friends Features

Twitter actively fosters interconnectivity. Users are linked by following each other. Content travels from person to person via retweets. Tweets themselves address specific users via mentions. We consider four types of friends (contacts): retweeting, mentioning, being retweeted, and being mentioned users. For each group separately, we extract features about language use, local time, popularity, etc. (cf. Table 6.1). Note that, due to limits of Twitter's REST API, we do not use the follower/followed relation beyond the total number of each as mentioned in the previous section.

6.2.1.3 Network Features

The network structure carries crucial information for the characterization of different types of communication. In fact, the usage of network features significantly helps in tasks like

political astroturf detection [240]. Our system reconstructs three types of networks: retweet, mention, and hashtag co-occurrence networks. Retweet and mention networks have users as nodes, with a directed link between a pair of users that follows the direction of information spreading: toward the user retweeting or being mentioned. Hashtag co-occurrence networks have undirected links between hashtag nodes when two hashtags occur together in a tweet. All networks are weighted according to the frequency of interactions or co-occurrences. For each network we compute a set of features, including in- and out-strength (weighted degree) distributions, density, clustering (cf. Table 6.1).

6.2.1.4 Time Features

Prior research suggests that the temporal signature of content production and consumption may reveal important information about online campaigns and their evolution [117,129,283]. To extract this signal we measure several temporal features of user activity. The most basic of these metrics is the rate of tweet production over various time periods. In addition, we capture the distribution of time intervals between events (cf. Table 6.1).

6.2.1.5 Content and Language Features

Many recent papers have demonstrated the importance of content and language features in revealing the nature of social media conversations [47, 90, 92, 184, 197, 209]. For example, deceiving messages generally exhibit informal language and short sentences [50]. Our system collects statistics about length and entropy of tweet text. Additionally, we extract language features by applying the *Part-of-Speech* (POS) tagging technique, which identifies different types of natural language components, or *POS tags*.[5] Tweets are therefore analyzed to study how POS tags are distributed (cf. Table 6.1).

[5]See: `www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html`

6.2.1.6 Sentiment Features

Sentiment analysis is a powerful tool to describe the emotions conveyed by a piece of text, and more broadly the attitude or mood of an entire conversation. Sentiment extracted from social media conversations has been used to forecast offline events including financial market fluctuations [42], and is known to affect information spreading [119, 207]. Our framework leverages several sentiment extraction techniques to generate various sentiment features, including *arousal*, *valence* and *dominance* scores [295], *happiness* score [172], *polarization* and *strength* [305], and *emotion* score [4] (cf. Table 6.1).

6.2.2 Model Evaluation

To train our system we initially used a publicly available labeled dataset consisting of 15K manually verified social bots and 16K legitimate (human) accounts identified via a *honeypot* approach [180]. We collected the most recent tweets produced by those accounts using the Twitter Search API.[6] We limited our collection to 200 public tweets from a user timeline and up to 100 of the most recent public tweets mentioning a user. This procedure yielded a dataset of 2.6 million tweets produced by manually verified social bots and 3 million tweets produced by users labeled as human. This bootstrap dataset helped us evaluate and compare the performance of several machine learning algorithms and the contributions of different feature sets.

We benchmarked our system using several off-the-shelf algorithms provided in the *scikit-learn* library [232]. In a generic evaluation experiment, the classifier under examination is provided with numerical vectors, each describing the features of an account. The classifier returns a numerical score in the unit interval. A higher score indicates a stronger likelihood that the account is a bot. A model's accuracy is evaluated by measuring the Area Under the

[6] `dev.twitter.com/rest/public`

Table 6.1: List of 1150 features extracted by our framework.

User meta-data	Screen name length	**Sentiment**	(***) Happiness scores of aggregated tweets
	Number of digits in screen name		(***) Valence scores of aggregated tweets
	User name length		(***) Arousal scores of aggregated tweets
	Time offset (sec.)		(***) Dominance scores of single tweets
	Default profile (binary)		(*) Happiness score of single tweets
	Default picture (binary)		(*) Valence score of single tweets
	Account age (days)		(*) Arousal score of single tweets
	Number of unique profile descriptions		(*) Dominance score of single tweets
	(*) Profile description lengths		(*) Polarization score of single tweets
	(*) Number of friends distribution		(*) Entropy of polarization scores of single tweets
	(*) Number of followers distribution		(*) Pos. emoticons entropy of single tweets
	(*) Number of favorites distribution		(*) Neg. emoticons entropy of single tweets
	Number of friends (S/R and rel. change)		(*) Emoticons entropy of single tweets
	Number of followers (S/R and rel. change)		(*) Pos. and neg. score ratio of single tweets
	Number of favorites (S/R and rel. change)		(*) Number of pos. emoticons in single tweets
	Number of tweets (per hour and total)		(*) Number of neg. emoticons in single tweets
	Number of retweets (per hour and total)		(*) Total number of emoticons in single tweets
	Number of mentions (per hour and total)		Ratio of tweets that contain emoticons
	Number of replies (per hour and total)		
	Number of retweeted (per hour and total)		
Friends (†)	Number of distinct languages	**Network (‡)**	Number of nodes
	Entropy of language use		Number of edges (also for reciprocal)
	(*) Account age distribution		(*) Strength distribution
	(*) Time offset distribution		(*) In-strength distribution
	(*) Number of friends distribution		(*) Out-strength distribution
	(*) Number of followers distribution		Network density (also for reciprocal)
	(*) Number of tweets distribution		(*) Clustering coeff. (also for reciprocal)
	(*) Description length distribution		
	Fraction of users with default profile		
	Fraction of users with default picture		
Content	(*,**) Frequency of POS tags in a tweet	**Timing**	(*) Time between two consecutive tweets
	(*,**) Proportion of POS tags in a tweet		(*) Time between two consecutive retweets
	(*) Number of words in a tweet		(*) Time between two consecutive mentions
	(*) Entropy of words in a tweet		

† We consider four types of connected users: retweeting, mentioning, retweeted, and mentioned.

‡ We consider three types of network: retweet, mention, and hashtag co-occurrence networks.

* Distribution types. For each distribution, the following eight statistics are computed and used as individual features: min, max, median, mean, std. deviation, skewness, kurtosis, and entropy.

** Part-of-Speech (POS) tag. There are nine POS tags: verbs, nuns, adjectives, modal auxiliaries, pre-determiners, interjections, adverbs, wh-, and pronouns.

*** For each feature, we compute mean and std. deviation of the weighted average across words in the lexicon.

receiver operating characteristic Curve (AUC) with 5-fold cross validation, and computing the average AUC score across the folds, as shown in Fig. 6.5. The best classification performance of 0.95 AUC was obtained by the *Random Forest* algorithm. In the rest of the paper we use the Random Forest model trained using 100 estimators and the Gini coefficient as a

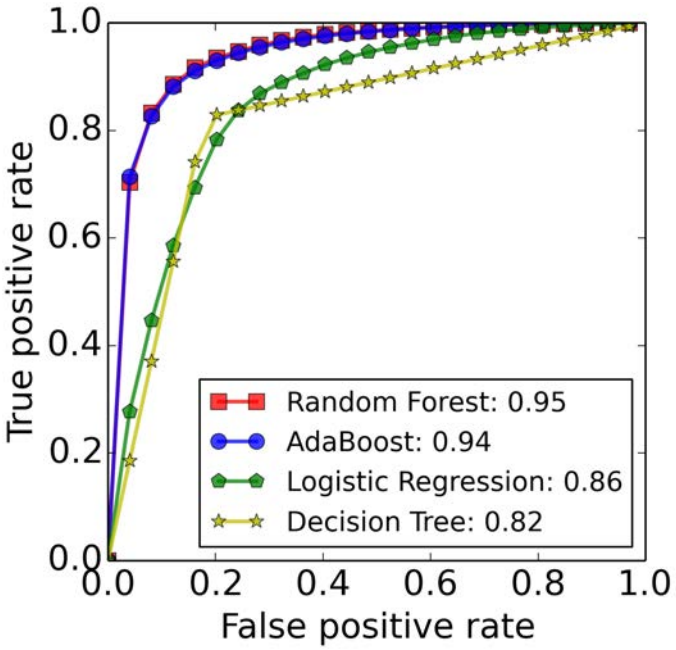

Figure 6.5: Classification performance of our system for four different classifiers. Accuracy is computed by five-fold cross validation and measured by the area under the ROC curve.

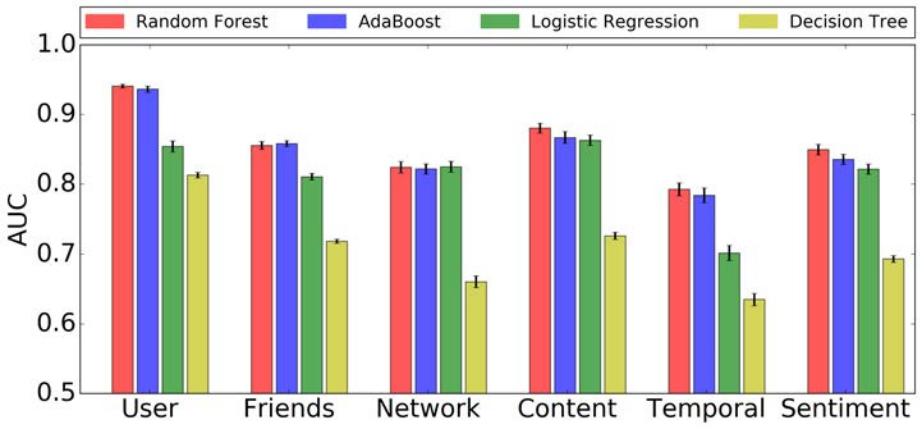

Figure 6.6: Performance across feature classes.

criterion of measuring the quality of splits.

To analyze the importance of each feature class, we can train the classifier using only the corresponding subset of features. We repeated the performance evaluation experiments considering only user, friends, network, content, temporal, and sentiment feature classes. In Fig. 6.6 we present the performance of classifiers using the different feature subsets. We achieved best performance with user meta-data features; content features were also shown

to be effective for the classification of social bots. Other feature classes yielded acceptable performance above 0.8 AUC.

6.3 Online Human-Bot Interactions: Detection, Estimation, and Characterization

In this section, we use our bot detection system to evaluate a large-scale collection of users. The performance of our detection system is evaluated against both an existing public dataset and an additional sample of manually annotated Twitter accounts. We enrich the models trained using existing bot data with the new annotations and investigate the effects of different datasets and classification models. We also classify a sample comprising millions of English-speaking active users. We use different models to estimate the percentage of Twitter accounts exhibiting social bot characteristics.

6.3.1 Model Improvement Using Manually Annotated Data

To obtain an updated evaluation of the accuracy of our classifier, we constructed an additional, manually annotated collection of Twitter user accounts. We leveraged these manual annotations to evaluate the model trained using the honeypot dataset and then to update the classifier's training data, producing a *merged* dataset to train a new model with better generalization for more sophisticated accounts.

6.3.1.1 Data Collection

Our data collection focused on *active* users producing content in English, as inferred from profile meta-data. We identified active users by monitoring a large Twitter stream, accounting for approximately 10% of public tweets, for 3 months starting in October 2015. Sampling from the public stream allows us to focus on active users while avoiding the biases of other

methods such as snowball and breadth-first sampling [130], which rely on the selection of an initial groups of users.

To restrict our sample to recently active users, we introduce the further criteria that they must have produced at least 200 tweets in total and 90 tweets during the three-month observation window (one per day on average). From our original sample, 14 million user accounts meet both criteria.We consider users in the highlighted area as active users. For each of these accounts, we collected their tweets through the Twitter Search API.[7] We restricted the collection to the most recent 200 tweets and 100 mentions of each user, as described earlier. Owing to Twitter API limits, this greatly improved our data collection speed. However this limitation adds noise to the features, due to the scarcity of data available to compute them.

6.3.1.2 Manual Annotations

We computed classification scores (defined in the unit interval) for each of the active accounts using our initial classifier trained on the honeypot dataset. We then grouped accounts by their bot scores, allowing us to evaluate our system across the spectrum of human and bot accounts. We randomly sampled 300 accounts from each bot score decile, yielding a balanced set of 3000 accounts. These were manually annotated by inspecting their public Twitter profile pages. In some cases there are obvious flags about bots, such as when an account uses a stock profile image or retweets every message of another account within seconds. In general, however, there is no simple set of rules to assess whether an account is human or bot. Each annotator analyzed profile appearance, content produced and retweeted, and interactions with other users in terms of retweets and mentions. The final decision reflects each annotator's opinion and are restricted to: human, social bot, or undecided. Accounts labeled as undecided were eliminated from further analysis.

[7] `http://dev.twitter.com/rest/public/search`

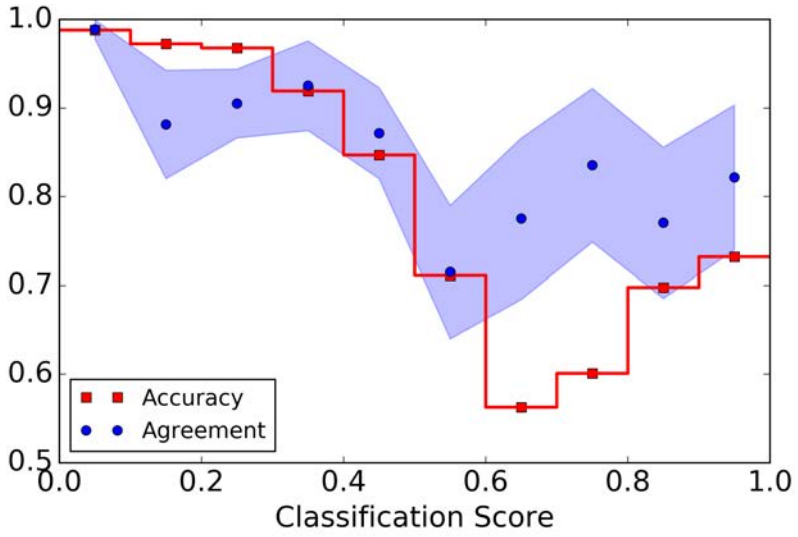

Figure 6.7: Accuracy of the model using the human annotations as the ground truth. Agreement is the average pairwise agreement of human annotators, presented with standard errors.

One author and four volunteers, familiar with Twitter, annotated all 3000 accounts. Each annotator was assigned a random sample of accounts from each decile. We enforced a minimum 10% overlap between annotations to assess the reliability of each annotator. This yielded an average pairwise agreement of 75% and moderately high inter-annotator agreement (Cohen's $\kappa = 0.41$). We also computed the agreement between annotators and classifier outcomes, assuming that a classification score above 0.5 is interpreted as a bot. This resulted in an average pairwise agreement of 79% and moderate $\kappa = 0.5$.

6.3.2 Evaluating Models Using Annotated Data

To evaluate our classification system trained on the honeypot dataset, we examined the classification accuracy separately for each bot-score decile. In Fig. 6.7, we present the accuracies of the model and inter-annotator agreements for annotated accounts in each bin. We achieved classification accuracy greater than 0.9 for the accounts in the $(0.0, 0.4)$ range, which includes mostly human accounts. We also observe accuracy above 0.7 for scores in the $(0.8, 1.0)$ range (mostly bots). Accuracy for boundary accounts ranges between 0.6 and 0.8.

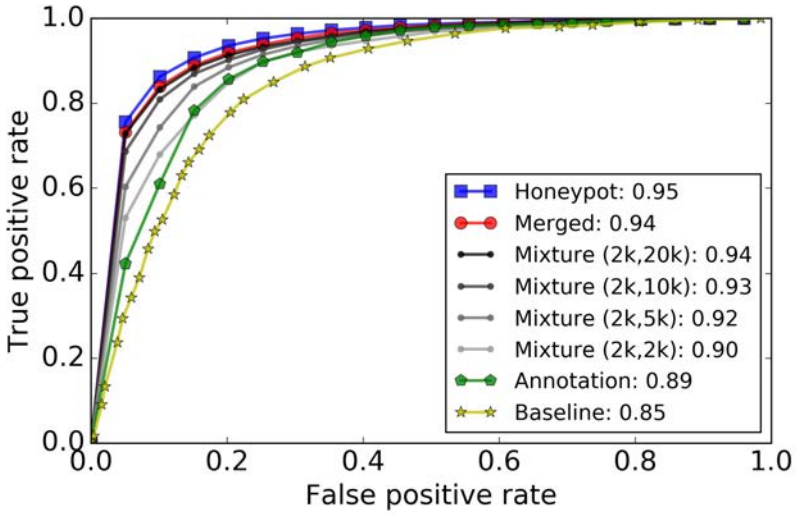

Figure 6.8: ROC curves of models trained and tested on different datasets. Accuracy is measured by AUC.

Intuitively, this range contains the most challenging accounts to label, making it difficult both for human annotators and for machine learning to achieve very high accuracy. When the accuracy of each bin is weighted by the population density in the large-scale sample, we obtain 86% accuracy overall.

We also compare annotator agreement scores for the accounts in each bot-score decile. We observe that agreement scores are higher for bins containing human accounts and lower for bots, indicating that it is more difficult for human annotators to identify social bots.

6.3.3 Dataset Effect on Model Accuracy

We can update our classification models by combining the manually annotated accounts with the honeypot dataset. We hypothesize that the recently collected bots in the annotated dataset may be more sophisticated than the ones obtained years earlier with the honeypot method. Fig. 6.8 illustrates the results of experiments designed to investigate our capability to detect such bots. The baseline ROC curve is obtained by testing the honeypot model, described in Sec. 6.2.2, on the manually annotated dataset. Unsurprisingly the baseline

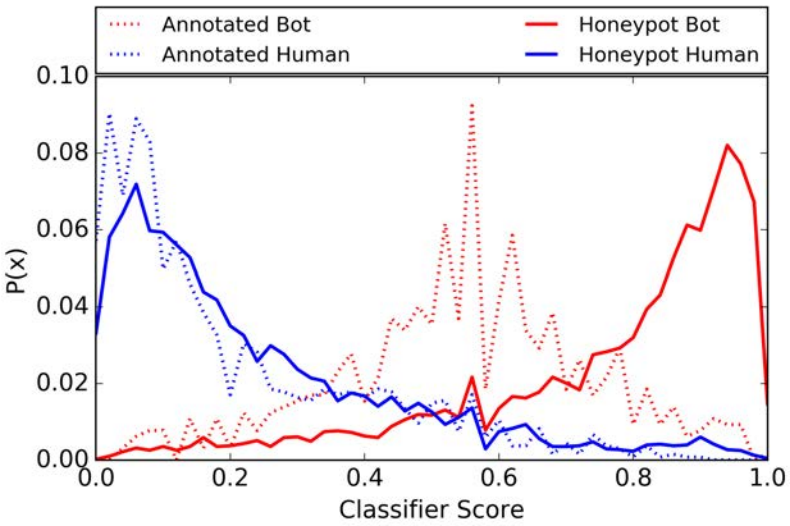

Figure 6.9: Distribution of classifier score for human and bot accounts in the two datasets.

accuracy (0.85 AUC) is lower than that obtained testing on the honeypot data (0.95 AUC), because the model is not trained on the newer bots. We created multiple balanced datasets and performed 5-fold cross-validation to evaluate the accuracy of the corresponding models:

- **Annotation**: We trained this model by only using annotated accounts and labels assigned by the majority of annotators. Our framework yields 0.89 AUC, a reasonable accuracy considering that the dataset contains recent and possibly sophisticated bots.
- **Merged**: We merged the honeypot and annotation datasets. The resulting classifier achieves 0.94 AUC, only slightly worse than the honeypot model although the dataset contains a variety of more recent and possibly sophisticated bots.
- **Mixture**: Using mixtures with different ratios of accounts from the annotated and honeypot datasets, we obtain an accuracy ranging between 0.90 and 0.94.

In Fig 6.9, we plot the distributions of classification scores for human and bot accounts according to each dataset. The mixture model trained on 2K annotated and 10K honeypot accounts is used to compute the scores. Human accounts in both datasets have similar distributions, peaked around 0.1. The difference between bots in the two datasets is more prominent. Classifiers produce larger scores peaked around 0.9 for the simpler bots in the

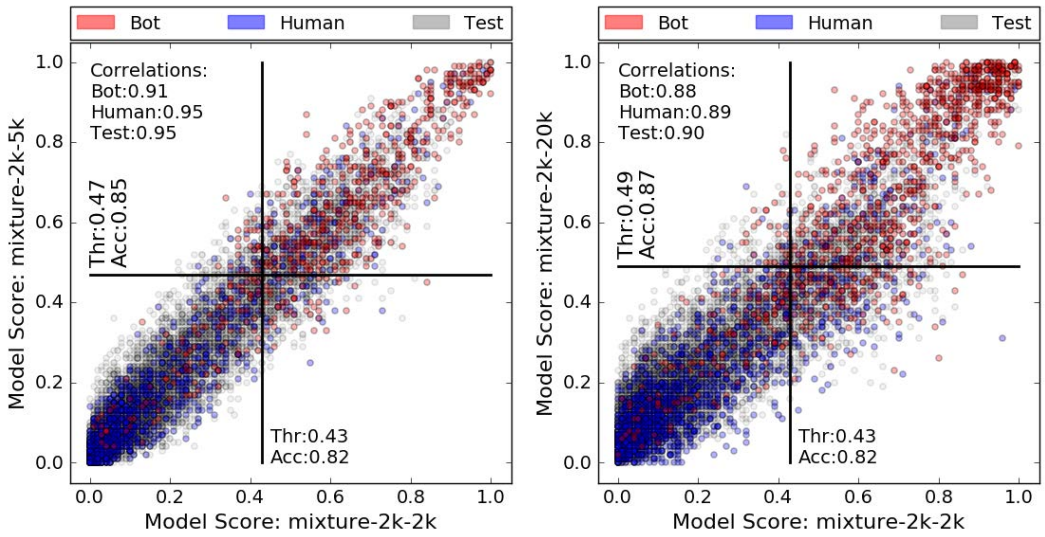

Figure 6.10: Comparison of prediction scores for different models. Each account is represented as a point in the scatter plot with a color determined by its ground-truth label. Additional test points are randomly sampled from our large-scale collection. Pearson correlations between scores are also reported, along with estimated thresholds and corresponding accuracies.

honeypot dataset; the newer bots have smaller scores, peaked around 0.6, supporting our hypothesis that they are more sophisticated, exhibiting some characteristics more similar to human behavior yielding lower scores on average. This distinction emphasises the importance of setting proper boundaries between human and bot accounts to discriminate them accurately.

We compared predicted scores by pairs of models for labeled human, bot and a random subset of users (see Fig. 6.10). As expected, both models assign lower scores for humans and higher for bots. High correlation coefficients indicate agreement between the models. To infer a suitable threshold in classification score that separates human and bot accounts for a given model, we computed classification accuracies for varying thresholds considering all accounts scoring below each threshold as human.

6.3.4 Estimation of Bot Population

In a 2014 report by Twitter to the US Securities and Exchange Commission, the company put forth an estimate that 8.5% of their user base consists of bots.[8] We would like to offer our own assessment of the proportion of bot accounts as measured with our approach. Since our framework provides a continuous bot score as opposed to a discrete bot/human judgement, we must first obtain an estimate of the bot-score threshold separating human and bot accounts to estimate the proportion of bot accounts.

We computed estimations for the population of social bots using different models. This approach allows us to identify lower and upper bounds for the prevalence of social bots. Models trained using the annotated dataset alone yield estimates of up to 15% of accounts being social bots. Recall that the honeypot dataset was obtained in 2011 and therefore does not include newer, more sophisticated bots. Thus models trained on the honeypot data alone are less sensitive to these sophisticated bots, yielding a more conservative estimate of 9%. Mixing the training data from these two sources results in estimates between these bounds depending on the ratio of the mixture, as illustrated in Fig. 6.11. Taken together, these numbers suggest that estimates about the prevalence of social bots are highly dependent on the definition and sophistication of the bots.

Some other remarks are in order: first, we do not exclude the possibility that very sophisticated bots exist that can systematically escape a human annotator's judgement. These complex bots may be active on Twitter, and therefore present in our datasets, and may have been incorrectly labeled as humans, making even the 15% figure a conservative estimate. Second, increasing anecdotal evidence suggests the presence on social media of hybrid human-bot accounts (sometimes referred to as *cyborgs*) that perform a broad range of automated actions with some human supervision and intervention [70, 77]. Some have

[8] `www.sec.gov/Archives/edgar/data/1418091/000156459014003474/twtr-10q_20140630.htm`

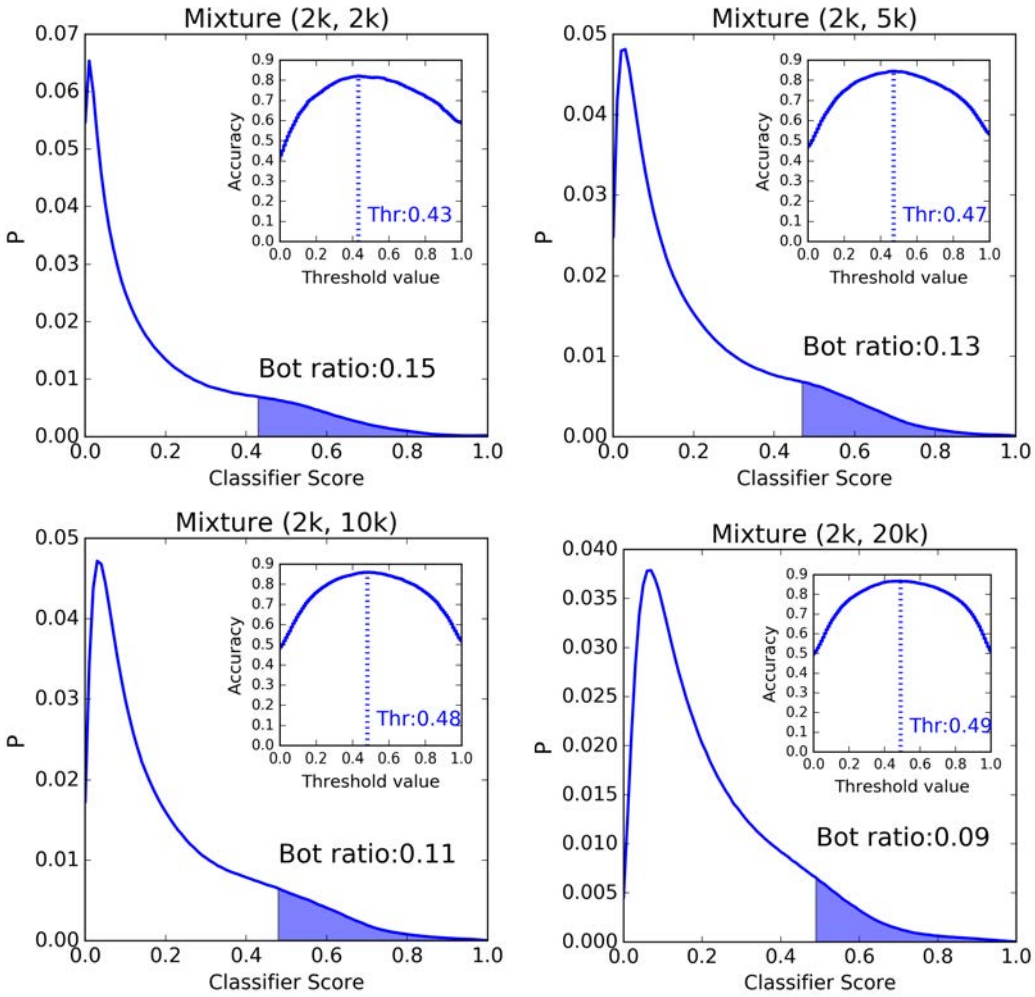

Figure 6.11: Estimation of bot population obtained from models with different sensitivity to sophisticated bots. The main charts show the score distributions based on our dataset of 14M users; accounts identified as bots are highlighted. The inset plots show how the thresholds are computed by maximizing accuracy. The titles of each subplot reflect the number of accounts from the annotated and honeypot datasets, respectively.

been allegedly used for terrorist propaganda and recruitment purposes. It remains unclear how these accounts should be labeled, and how pervasive they are.

6.3.5 Characterization of User Interactions

Let us next characterize social connectivity, information flow, and shared properties of users. We analyze the creation of social ties by accounts with different bot scores, and their interactions through shared content. We also cluster accounts and investigate shared properties

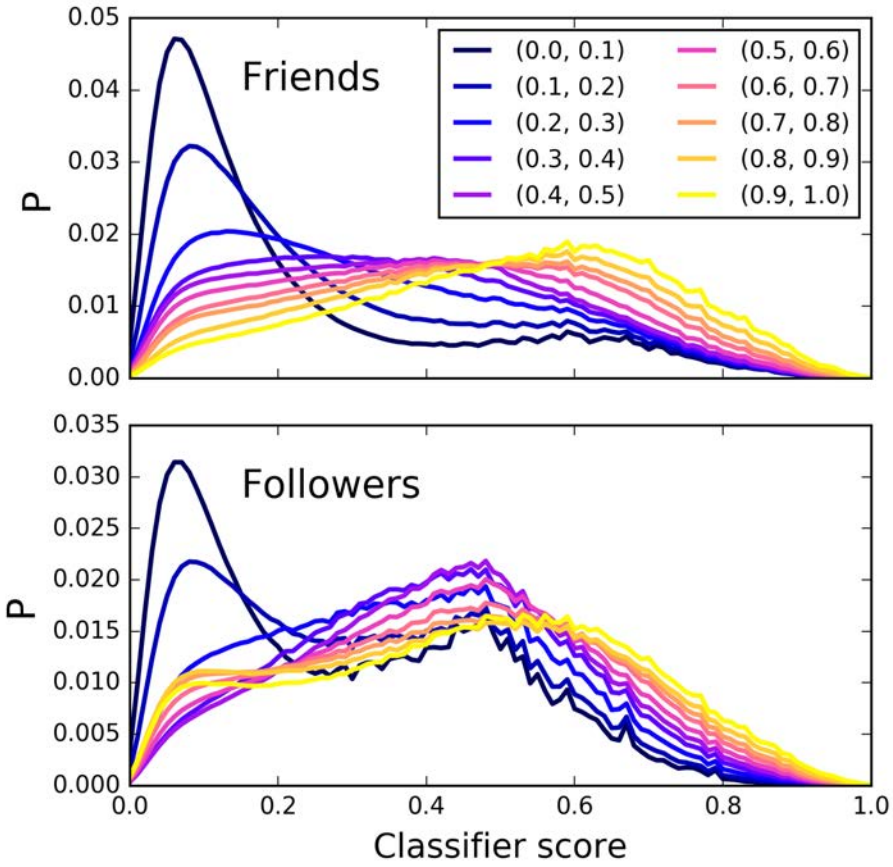

Figure 6.12: Distributions of bot scores for friends (top) and followers (bottom) of accounts in different score intervals.

of users in each cluster. Here and in the remainder of this paper, bot scores are computed with a model trained on the *merged* dataset.

6.3.5.1 Social Connectivity

To characterize the social connectivity, we collected the social networks of the accounts in our dataset using the Twitter API. Resulting friend and follower relations account for 46 billion social ties, 7 billion of which represent ties between the initially collected user set.

Our observations on social connectivity are presented in Fig. 6.12. We computed bot-score distributions of friends and followers of accounts for each score interval. The dark line in the top panel shows that human accounts (low score) mostly follow other human accounts.

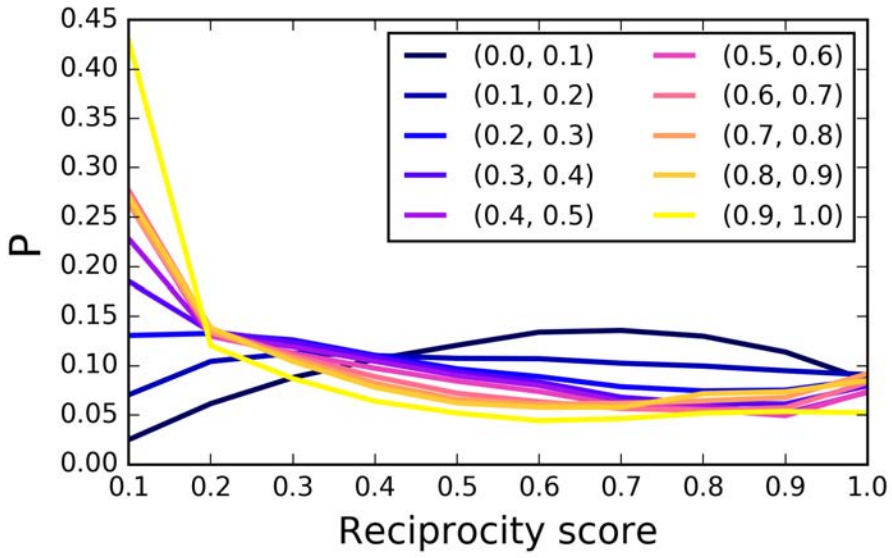

Figure 6.13: Distribution of reciprocity scores for accounts in different score intervals.

The dark line in the bottom panel shows a principal peak around 0.1 and a secondary one around 0.5. This indicates that humans are typically followed by other humans, but also by sophisticated bots (intermediate scores). The lines corresponding to high scores in the two panels show that bots tend to follow other bots and they are mostly followed by bots. However simple bots (0.8–1.0 ranges) can also attract human attention. This happens when, e.g., humans follow benign bots such as those that share news. This gives rise to the secondary peak of the red line in the bottom panel. In summary, the creation of social ties leads to a homophily effect.

Fig. 6.13 illustrates the extent to which connections are reciprocated, given the nature of the accounts forming the ties. The *reciprocity score* of a user is defined as the fraction of friends who are also followers. We observe that human accounts reciprocate more (dark line). Increasing bot scores correlate with lower reciprocity. We also observe that simple bot accounts (0.8–1.0 ranges) have bimodal reciprocity distributions, indicating the existence of two distinct behaviors. The majority of high-score accounts have reciprocity score smaller than 0.2, possibly because simple bots follow users at random. The slight increase as the

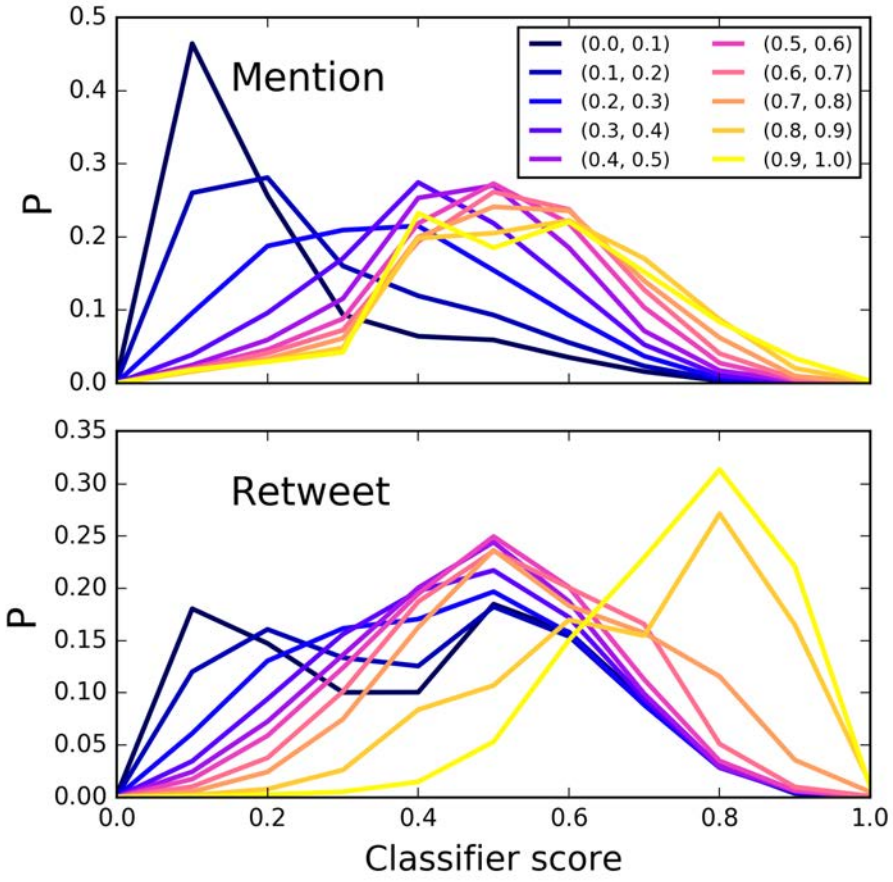

Figure 6.14: Bot score distributions of users mentioned (top) and retweeted (bottom) by accounts with different scores.

reciprocity score approaches one may be due to botnet accounts that coordinate by following each other.

6.3.5.2 Information Flow

Twitter is a platform that fosters social connectivity and the broadcasting of popular content. In Fig. 6.14 we analyze information flow in terms of mentions/retweets as a function of the score of the account being mentioned or retweeted.

Simple bots tend to retweet each other (lines for scores in the 0.8–1.0 ranges peak around 0.8 in the bottom panel), while they frequently mention sophisticated bots (peaking around 0.5 in the top panel). More sophisticated bots (scores in the 0.5–0.7 ranges) retweet, but

do not mention humans. They might be unable to engage in meaningful exchanges with humans. While humans also retweet bots, as they may post interesting content (see peaks of the dark lines in the bottom panel), they have no interest in mentioning bots directly (dark lines in the top panel).

6.3.5.3 Clustering Accounts

To characterize different account types, let us group accounts into behavioral clusters. We apply K-Means to normalized vectors of the 100 most important features selected by our Random Forests model. We identify 10 distinct clusters based on different evaluation criteria, such as silhouette scores and percentage of variance explained. In Fig 6.15, we present a 2-dimensional projection of users obtained by a dimensionality reduction technique called t-SNE [192]. In this method, the similarity between users is computed based on their 100-dimensional representation in the feature space. Similar users are projected into nearby points and dissimilar users are kept distant from each other.

Let us investigate shared cluster properties by manual inspection of random subsets of accounts from each cluster. Three of the clusters, namely `C0`–`C2`, have high average bot scores. The presence of significant amounts of bot accounts in these clusters was manually verified. These *bot* clusters exhibit some prominent properties: cluster `C0`, for example, consists of legit-looking accounts that are promoting themselves (recruiters, porn actresses, etc.). They are concentrated in the lower part of the 2-dimensional embedding, suggesting homogeneous patterns of behaviors. `C1` contains spam accounts that are very active but have few followers. Accounts in `C2` frequently use automated applications to share activity from other platforms like YouTube and Instagram, or post links to news articles. Some of the accounts in `C2` might belong to actual humans who are no longer active and their posts are mostly sent by connected apps.

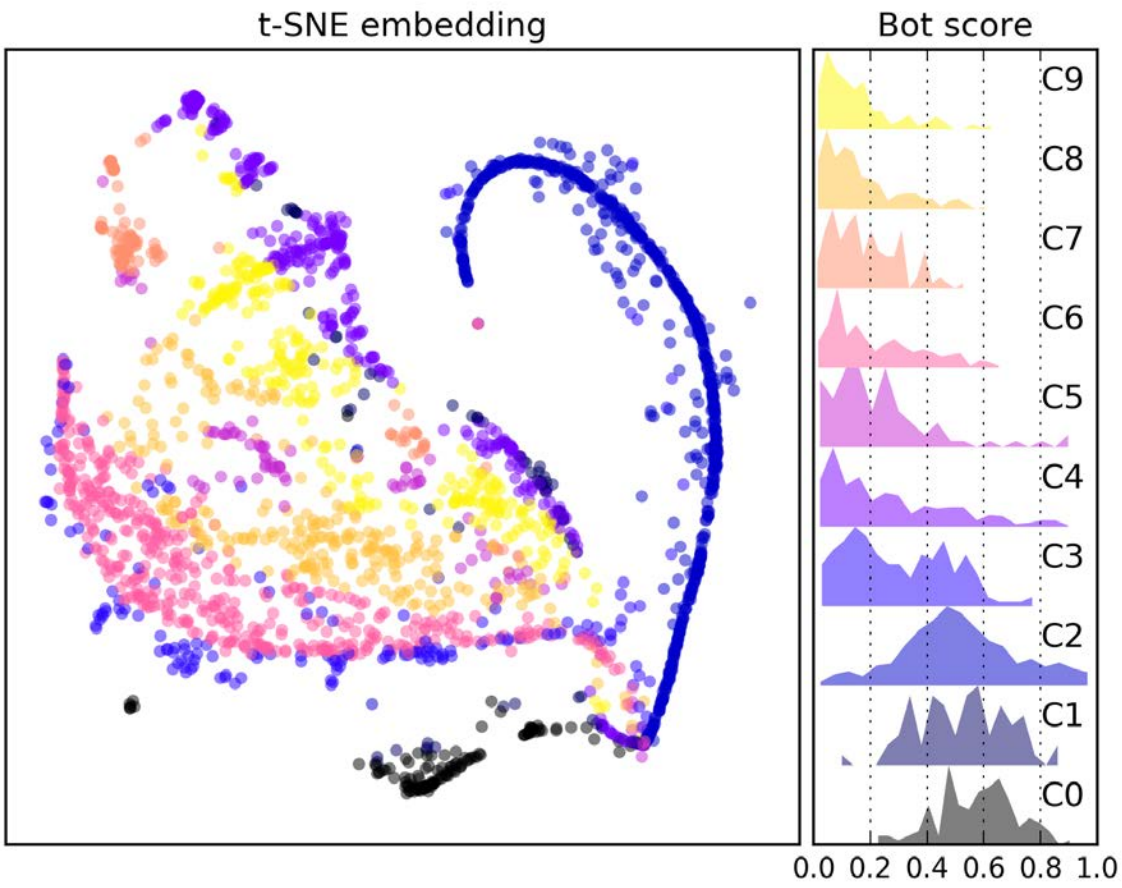

Figure 6.15: t-SNE embedding of accounts. Points are colored based on clustering in high-dimensional space. For each cluster, the distribution of scores is presented on the right.

Cluster `C3` contain a *mix* of sophisticated bots, cyborg-like accounts (mix of bot and human features), and human users. Clusters of predominantly *human* accounts, namely `C4`–`C9`, separate from one another in the embedding due to different activity styles, user popularity, content production and consumption patterns. For instance, accounts in `C7` engage more with their friends, unlike accounts from `C8` that mostly retweet with little other forms of interaction. Clusters `C5`, `C6`, and `C9` contain common Twitter users who produce experiential tweets, share pictures, and retweet their friends.

6.4 Conclusions

Social media make it easy for accounts controlled by hybrid or automated approaches to create content and interact with other accounts. Our project aims to identify these bots. Such a classification task could be a first step toward studying modes of communication among different classes of entities on social media.

In this article, we presented a framework for bot detection on Twitter. We introduced our machine learning system that extracts more than a thousand features in six different classes: users and friends meta-data, tweet content and sentiment, network patterns, and activity time series. We evaluated our framework when initially trained on an available dataset of bots. Our initial classifier achieves 0.95 AUC when evaluated by using 5-fold cross validation. Our analysis on the contributions of different feature classes suggests that user meta-data and content features are the two most valuable sources of data to detect simple bots.

To evaluate the performance of our classifier on a more recent and challenging sample of bots, we randomly selected Twitter accounts covering the whole spectrum of classification scores. The accuracy of our initial classifier trained on the honeypot dataset decreased to 0.85 AUC when tested on the more challenging dataset. By retraining the classifier with the two datasets merged, we achieved high accuracy (0.94 AUC) in detecting both simple and sophisticated bots.

We also estimated the fraction of bots in the active English-speaking population on Twitter. We classified nearly 14M accounts using our system and inferred the optimal threshold scores that separate human and bot accounts for several models with different mixes of simple and sophisticated bots. Training data have an important effect on classifier sensitivity. Our estimates for the bot population range between 9% and 15%. This points to the importance of tracking increasingly sophisticated bots, since deception and detection

technologies are in a never-ending arms race.

To characterize user interactions, we studied social connectivity and information flow between different user groups. We showed that selection of friends and followers are correlated with accounts bot-likelihood. We also highlighted how bots use different retweet and mention strategies when interacting with humans or other bots.

We concluded our analysis by characterizing subclasses of account behaviors. Clusters identified by this analysis point mainly to three types of bots. These results emphasize that Twitter hosts a variety of users with diverse behaviors; this is true for both human and bot accounts. In some cases, the boundary separating these two groups is not sharp and an account can exhibit characteristics of both.

CHAPTER 7

Conclusions

Social media are important tools and their efficient use promotes information dissemination, fosters connectivity between individuals, and helps accessibility and transparency of knowledge. In the Internet age, we are endowed with the capability to observe activities of the millions, to model interactions between individuals, and to discover unknown properties of society. Using techniques from network science, computer science, and social science, I presented studies of properties of socio-technical systems and built tools to ensure their robustness against malicious intentions.

This dissertation presents several studies on the topics of information diffusion, online discourse, and detection of campaigns and social bots. Our analysis on trend diffusion shows that geography still plays a significant role in information dissemination. Analysis on censorship reveals that withheld tweets foster curiosity of readers and yield increases in popularity of censored user and content. Similar effects have been observed in printed media, where editors leave censored content blank in their publications to draw attention to the removed content. Our campaign detection framework also points to the importance of user features to distinguish promoted content. I also built a system to detect social bots and analyze their interactions on Twitter.

7.1 Summary and Discussion of Contributions

In Chapter 2, I discuss the relationships between work presented in this dissertation and existing social science literature. The following sections summarize the contributions of Chapters 4, 5, and 6 and discuss their implications for future work.

7.1.1 Online Discourse

Online discourse is a broad topic that we discussed in this dissertation. I studied online discourse from different perspectives: geography of information diffusion, effects of censorship, and characterization of user roles during social protests. We learned lessons about how geography plays a role on local trends, but national trends and censorship rely on global ties. Our analysis of the Gezi movement highlights the importance of collective behavior and points to the interplay between external events and online activities. Here I summarize observations from our analysis of trend diffusion and online censorship.

- We describe a procedure to build a directed and weighted temporal dependence network to infer the trendsetting and trend-following relationships among locations. We provide a statistical characterization of trends, describing how they are distributed in space and time.
- We describe two different dynamics that govern popularity of trends at the country level, one for cities in each local geographic area and one for metropolitan areas. We conclude by highlighting that the major metropolitan areas shape the country trends significantly more than all other locations. We propose an interpretation for the trendsetting role of major metropolitan areas, by noting their correspondence with air traffic hubs and conjecturing that trends travel through air passengers, just as infectious diseases.

- We explore the spatio-temporal characteristics of censorship, that is, which governments requested censorship and how the volume of these requests changed over time.
- We show that IP-based censorship on Twitter is not an effective mechanism. We analyze language and timezone preferences of the retweeting users as a proxy to user location and show that the diffusion of the censored content is not limited to the boundaries of the requesting governments, but spans larger populations.
- We point to an important observation on how the amount of censorship correlates with the changes of user behavior. We observe an increase in the number of friends and attention paid to censored content for those users targeted by increasing censorship.

Our analysis of social protests in Turkey helps us understand the dynamics of social protests and user roles:

- We present methods to extract topically focused conversations about the social uprising surrounding Gezi Park and related trending topics of conversation on Twitter.
- We explore the spatio-temporal characteristics of the conversation; that is, where tweets about Gezi Park originated and what locations shared most similar topics and trends. This analysis yields clusters of cities that are mostly consistent with the country's geopolitics.
- We analyze the emerging characteristics of users involved in the conversation about Gezi Park protests on Twitter, the roles they played in this context, and how these roles evolved as the protest unfolded. We find that influence was redistributed in the user population over time, making the conversation more democratic.
- We show that online user behavior was affected by external factors, such as speeches by political leaders or police action to hinder or suffocate the protests.
- We focus on leveraging hand-coded data with automated techniques to identify distinct behavioral groups. We have learned that the primary role for users was in information

dissemination to other participants in the demonstrations, but few messages indicated any leadership role and users who tweeted directive messages did not do so consistently.

This section of the dissertation mainly addresses studies of information dissemination and geography. I studied the role of geography on information diffusion to show that geography still plays a role in information diffusion. One can study the multiplex structure of information and social networks to reveal hidden branches of diffusion trees. My analysis of censorship shows interesting parallels with the historic practices of censorship, where censorship promotes reader curiosity and motivates them to search for details.

We also studied social protest to investigate how different user roles emerge and evolve over time. Emergent behavior in social systems yields interesting group dynamics: polarization, marginalization, and social upheavals. To study socio-technical systems, it is crucial to understand group behaviors shaped by social norms. In future work, one can identify the latent factors shaping group membership and changes in personal position for a cause. Social media can influence and shape public opinion, and misinformation and social bots play a significant role in affecting belief systems. I also want to test how public discourse by politicians on social media can be used to direct limited attention to less important conversations. I want to study persuasion and deception in the age of limited attention.

7.1.2 Campaign Detection

Online discourse can be manipulated and controlled. To study online campaigns, we focus on trending memes on Twitter and on a special case of promotion, namely advertising, because it provides a convenient operational definition of social media campaign. We formally define the task of discriminating between organic and promoted trending memes. We built a machine learning framework that exploits hundreds of social media signals over time to capture signatures of orchestrated campaigns. This approach takes a first step toward the

development of computational methods for the early detection of information campaigns. We make the following contributions:

- We explore different methods for encoding feature time series. Using millions of tweets containing trending hashtags, we achieve 75% AUC score for early detection, increasing to above 95% after trending.
- We studied the robustness of systems to random shifts on temporal signals and pointed to the strength of algorithms that capture patterns in time series.
- One of the advantages of our framework is that of providing interpretable feature classes. We explored how content, network, and user features affect detection performance. Extensive feature analysis revealed that signatures of campaigns can be detected early, especially by leveraging content and user features. After the trending point, network and temporal features become more useful.

The lessons learned from this project can be used to study complex persuasion. Advertisers and political campaign organizers are actively developing strategies to reach out and communicate with their targeted audience. Efforts in designing viral online campaigns yield tools for modern marketing strategies. Unfortunately, entities with malicious intentions can also benefit from such systems and adopt them to achieve their goals. Traditionally, successful campaigns rely on carefully designed messages and punctual timing. Experts in social psychology can identify possible concepts to frame campaigns for targeted groups. The abundance of digital data and developments in personalization might facilitate targeted campaigns and the ability to rapidly evaluate the effects of different strategies and frames.

In future work, one can build automated techniques to identify campaigns by detecting behavioral anomalies at the conversation, account, or tweet level. Observations of account activities at the group level can provide insightful details about coordination.

7.1.3 Social Bots

Increasing evidence suggests that social bots have become a major problem for communication systems. We propose a framework to extract a large collection of features from data and meta-data about social media users, including friends, tweet content and sentiment, network patterns, and activity time series. We use these features to train highly-accurate models to identify bots. For a generic user, we produce a $[0, 1]$ score representing the likelihood that the user is a bot. Our research on social bots was timely and we made important contributions in this area. We published a highly cited review paper on social bots and released an online bot detection system for academic and public use. Our research on social bots yielded the following contributions:

- We participated in the DARPA bot detection challenge and completed the task as the second fastest and third most accurate team.
- We classified a sample comprising millions of English-speaking active Twitter users. We used different models to infer thresholds in the bot score that best discriminate between humans and bots. We estimated that the percentage of Twitter accounts exhibiting social bot behaviors is between 9% and 15%.
- We characterized friendship ties and information flow between users that show behaviors of different nature: human and bot-like. Humans tend to interact with more human-like accounts than bot-like ones, on average. Reciprocity of friendship ties is higher for humans. Simple bots target users more or less randomly, while sophisticated bots can choose targets based on their intentions.
- Clustering analysis revealed certain specific behavioral groups of accounts. Manual investigation of samples extracted from each cluster points to three distinct bot groups: spammers, self promoters, and accounts that post content from connected applications.
- Our social bot classification framework was released online for public use. Since its

release date we received millions of requests.

The intents and strategies of malicious entities such as social bots and orchestrated campaigns are either fully automated by software or directed by motivated human agents. Armies of social bots and misinformation campaigns are executed to promote ideas, advertise products, or sway public opinion. We have been observing social bots that attempt to persuade, influence, and deceive. Recent advances in deep-learning technologies accelerate fake persona generation [38, 185] and conversation models for social bots [186, 262]. Such technologies make social bots difficult to detect and provide an advantage in this arms race.

In future work, I want to use my experience in the identification of social bots and early detection of campaigns to isolate those activities and study their strategies in depth. I am interested in building detection systems that can evolve to lead in this arms-race by exploring behavioral signatures of users and characterizing their strategies.

Another important direction that is worth taking is studying the phenomenon called *account recycling*. My guess is that active Twitter accounts remove all their content and start a new persona for different campaigns or agendas. In this way, these accounts appear to have a history and maintain their followers during the transition. This could be easily monitored by tweet deletion notices.

7.2 Other Areas of Future Work

I am excited about the opportunities to mine social signals for gaining new insights about human behavior and society. The world we have been experiencing is changing and we have more accurate data with higher temporal resolution, as well as reflecting a detailed picture of individual lives. The ethical collection of multi-modal data about individuals will be instrumental in understanding human behaviors. In the future, I want to develop new models and tools to study complexity in terms of analyzing behaviors of individuals. The

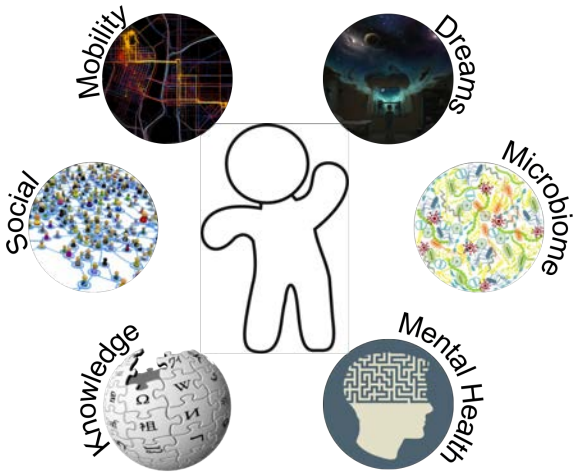

Figure 7.1: As a future work, expertise gained by studying mobility, social networks and knowledge networks can be applied to important problems around mental health, microbiome, and dream research.

most significant questions can be answered through connecting relevant problems that lie in organized complexity. Interdisciplinary research is crucial to connect existing knowledge to discover novel synthesis. I trained as a physicist and computer scientist that direct me to study in domains of EEG signal processing [280], modeling group behaviors using statistical mechanics, and analyzing fluctuations of proteins [286]. I always try to combine different school of thoughts when approaching a novel problems.

My long-term research goal is to develop models that describe dynamically changing intents and actions of individuals and groups. Network analysis, causal inference, and statistical learning techniques are core methods and I also want to employ deep learning models as predictive models in my research. Some of these models have advantages in terms of accuracy. However, certain sensitive domains require models with high interpretability and explanations of outcomes. Combinations of models and awareness of their limitations are important to study behaviors of individuals.

Modeling and detecting strategies employed by users is crucial for many reasons: understanding intents behind their actions, improving their well-being, and characterizing inter-

actions between groups. Deviations from the regular patterns can also point to important events and pre-cursors of significant transitions. Understanding changes in behavior helps to study mood changes and to identify significant life events. I believe that my research has potential implications for improving individual well-being, discovering new knowledge on how diseases and mental health problems progress, and understanding the nature of conflicts between groups. In the following, I describe several future directions I am excited to pursue (Fig. 7.1).

Identifying the intents of individuals and improving their well-being. One of the applications of ego-centric network research is to model mental health problems. In this domain, I would like to infer whether a user has issues like bipolar disorder and depression based on prior online interactions. To improve such inferences, I am studying the transfer of knowledge about users across platforms and datasets by employing deep learning, statistical learning, and causal inference. My goal is to build models for interconnected data sources to highlight the relationships between user attributes and behavioral markers. I am not only interested in studying social networks, communication, and mobility, but also health related precursor signals collected from biological data, personal logs, and other ego-centric measurements. Once a particular group of people is selected on one platform, users with similar characteristics can be identified on other platforms. Additional features about the group can be extracted from these platforms to improve the inference model and predict user behaviors. Developing ethical methods and tools is an important challenge of this project. I want to study how privacy of the users can be preserved while research efforts are devoted to learning about human behaviors. My goal is to formulate new hypotheses about disease progression and develop mechanisms for support.

Studying dreams to decipher the unconscious mind. I want to pursue a personal interest in dreams by building collaborations with clinical psychologists. Previously, I

worked on multi-cultural analysis of dream interpretations to highlight global archetypes and cultural differences [285]. Recently, I have been analyzing individual dream journals. Data driven research to understand the meaning of dreams and their implications on real life can be further improved by controlled experiments and data collection through mobile devices. Collaborative work in this area, in my opinion, will be greatly rewarding to understand unconscious behaviors.

BIBLIOGRAPHY

[1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206. ACM, 2015.

[2] 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.

[3] Sean Aday, Henry Farrel, Marc Lynch, John Sides, John Kelly, and Ethan Zuckerman. Blogs and bullets: New media in contentious politics. Technical report, U.S. Institute of Peace, 2010.

[4] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. ACL, 2011.

[5] C.C. Aggarwal and K. Subbian. Event detection in social streams. In *Proceedings of SIAM International Conference on Data Mining*, 2012.

[6] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. People are strange when you're a stranger: Impact and influence of bots on social networks. In *Proc. 6th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2012.

[7] Sadaf R Ali and Shahira Fahmy. Gatekeeping and citizen journalism: The use of social media during the recent uprisings in iran, egypt, and libya. *Media, War & Conflict*, 6(1):55–69, 2013.

[8] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[9] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[10] Lorenzo Alvisi, Allen Clement, Alessandro Epasto, Silvio Lattanzi, and Alessandro Panconesi. Sok: The evolution of sybil defense via social networks. In *Proc. IEEE Symposium on Security and Privacy (SP)*, pages 382–396, 2013.

[11] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[12] Tim Arango. Protests in turkey reveal a larger fight over identity. *NY Times*, June 2, 2013.

[13] Aristotle and George A Kennedy. *Rhetoric*. JSTOR, 1992.

[14] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

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

[16] Norman T.J. Bailey. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[17] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proc. 4th ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 65–74, 2011.

[18] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489, 2009.

[19] David Bamman, Brendan O'Connor, and Noah Smith. Censorship and deletion practices in chinese social media. *First Monday*, 17(3), 2012.

[20] Raquel A Baños, Javier Borge-Holthoefer, and Yamir Moreno. The role of hidden influentials in the diffusion of online information cascades. *EPJ Data Science*, 2(1):1–16, 2013.

[21] Raquel A Baños, Javier Borge-Holthoefer, Ning Wang, Yamir Moreno, and Sandra González-Bailón. Diffusion dynamics with changing network composition. *Entropy*, 15(11):4553–4568, 2013.

[22] Frank M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5), 1969.

[23] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[24] Roy L Behr and Shanto Iyengar. Television news, real-world cues, and changes in the public agenda. *Public Opinion Quarterly*, 49(1):38–57, 1985.

[25] Amy Bellmore, Angela J Calvin, Jun-Ming Xu, and Xiaojin Zhu. The five w's of "bullying" on twitter: who, what, why, where, and when. *Computers in human behavior*, 44:305–314, 2015.

[26] R.D. Benford. An insider's critique of the social movement framing perspective. *Sociological Inquiry*, 67(4):409–430, 1997.

[27] R.D. Benford and D.A. Snow. Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26(1):611–639, 2000.

[28] W. Bennett. Communicating global activism: Strength and vulnerabilities of networked politics. *Information, Communication & Society*, 6(2):143–168, 2003.

[29] W.L. Bennett. Changing citizenship in the digital age. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, pages 1–24, 2007.

[30] William L Benoit. *Seeing spots: A functional analysis of presidential television advertisements, 1952-1996*. Greenwood Publishing Group, 1999.

[31] J Berger and Jonathan Morgan. The isis twitter census: Defining and describing the population of isis supporters on twitter. *The Brookings Project on US Relations with the Islamic World*, 3:20, 2015.

[32] Edward L Bernays. *Crystallizing public opinion*. Open Road Media, 2015.

[33] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proc. of AAAI Workshop on Knowledge Discovery in Databases*, pages 359–370. Seattle, WA, 1994.

[34] 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):e0118093, 02 2015.

[35] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11), 2016.

[36] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 355–356. ACM, 2015.

[37] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proc. 22nd Intl. Conf. on World Wide Web (WWW)*, pages 119–130, 2013.

[38] Parminder Bhatia, Marsal Gavalda, and Arash Einolghozati. soc2seq: Social embedding meets conversation model. *arXiv preprint arXiv:1702.05512*, 2017.

[39] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[40] 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.

[41] Menahem Blondheim. *News over the wires: The telegraph and the flow of public information in America, 1844-1897*. Number 42. Harvard University Press, 1994.

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

[43] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*, 2016.

[44] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI 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, 2012.

[45] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz, et al. Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. *PLoS One*, 6(8):e23883, 2011.

[46] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proc. 27th Annual Computer Security Applications Conf.*, 2011.

[47] Federico Botta, Helen Susannah Moat, and Tobias Preis. Quantifying crowd size with mobile phone and twitter data. *Royal Society open science*, 2(5):150162, 2015.

[48] danah boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. 2007. *Journal of Computer-Mediated Communication*, 13(1), 2010.

[49] Ray Bradbury. *Fahrenheit 451: A Novel*. Simon and Schuster, 2012.

[50] E Briscoe, S Appling, and H Hayes. Cues to deception in social media communications. In *Proc. Hawaii Intl. Conf. on System Sciences*, 2014.

[51] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[52] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment*, 4(10):646–656, 2011.

[53] Ceren Budak and Duncan J Watts. Dissecting the spirit of gezi: Influence vs. selection in the occupy gezi movement. *Selection in the Occupy Gezi Movement.(May 25, 2015)*, 2015.

[54] Sam Burnett and Nick Feamster. Making sense of internet censorship: a new frontier for internet measurement. *ACM SIGCOMM Computer Communication Review*, 43(3):84–89, 2013.

[55] Alison M Buttenheim, Karthik Sethuraman, Saad B Omer, Alexandra L Hanlon, Michael Z Levy, and Daniel Salmon. Mmr vaccination status of children exempted from school-entry immunization mandates. *Vaccine*, 33(46):6250–6256, 2015.

[56] J. Byrne. Occupy the media: Journalism for (and by) the 99 percent. In Janet Bryne, editor, *The Occupy Handbook*, pages 256–264. Little, Brown, 2012.

[57] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 197–210, 2012.

[58] Carlos Carvalho, Nicholas Klagge, and Emanuel Moench. The persistent effects of a false news shock. *Journal of Empirical Finance*, 18(4):597–615, 2011.

[59] Manuel Castells. Communication, power and counter-power in the network society. *International journal of communication*, 1(1):29, 2007.

[60] 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.

[61] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329:1194–1197, 2010.

[62] Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

[63] Shelly Chaiken, Wendy Wood, and Alice H. Eagly. Principles of persuasion. 1996.

[64] Julian C Chambliss. Superhero comics: Artifacts of the us experience. *Juniata Voices*, 12:149, 2012.

[65] 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*, pages 925–936. ACM, 2014.

[66] Justin Cheng, Lada A Adamic, Jon M Kleinberg, and Jure Leskovec. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web*, pages 671–681. International World Wide Web Conferences Steering Committee, 2016.

[67] N. Chomsky. *Occupy*. Zuccotti Park Press, 2012.

[68] A. Choudhary, W. Hendrix, K. Lee, D. Palsetia, and W.K. Liao. Social media evolution of the egyptian revolution. *Communications of the ACM*, 55(5):74–80, 2012.

[69] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proc. 26th annual computer security applications conference*, pages 21–30, 2010.

[70] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Tran. Dependable and Secure Computing*, 9(6):811–824, 2012.

[71] Robert B. Cialdini. Influence: The psychology of persuasion. 1993.

[72] Robert B Cialdini. Science and practice. 2001.

[73] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. The production of information in the attention economy. *Scientific reports*, 5, 2015.

[74] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PLoS ONE*, 10(6):e0128193, 06 2015.

[75] Fabio Ciulla, Delia Mocanu, Andrea Baronchelli, Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Beating the news using social media: the case study of American Idol. *EPJ Data Science*, 1(1):1–11, 2012.

[76] Eric M Clark, Chris A Jones, Jake Ryland Williams, Allison N Kurti, Michell Craig Nortotsky, Christopher M Danforth, and Peter Sheridan Dodds. Vaporous marketing: Uncovering pervasive electronic cigarette advertisements on twitter. *arXiv preprint arXiv:1508.01843*, 2015.

[77] Eric M Clark, Jake Ryland Williams, Chris A Jones, Richard A Galbraith, Christopher M Danforth, and Peter Sheridan Dodds. Sifting robotic from organic text: a natural language approach for detecting automation on twitter. *Journal of Computational Science*, 16:1–7, 2016.

[78] Steve Clarke. Conspiracy theories and conspiracy theorizing. *Philosophy of the Social Sciences*, 32(2):131–150, 2002.

[79] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020, 2006.

[80] Ross F Collins. A battle for humor: satire and censorship in le bayard. *Journalism & Mass Communication Quarterly*, 73(3):645–656, 1996.

[81] British Parliamentary Recruiting Committee. Daddy, what did you do in the great war?, 1915. [Online; accessed February 27, 2017].

[82] M.D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on Twitter. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[83] 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:e55957, 2013.

[84] Michael D Conover, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. The digital evolution of Occupy Wall Street. *PloS ONE*, 8:e64679, 2013.

[85] Rion Brattig Correia, Lang Li, and Luis M Rocha. Monitoring potential drug interactions and reactions via network analysis of instagram user timelines. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 21, page 492. NIH Public Access, 2016.

[86] Thomas M Cover and Peter E Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[87] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17:124–147, 2013.

[88] Alberto Dainotti, Claudio Squarcella, Emile Aben, Kimberly C Claffy, Marco Chiesa, Michele Russo, and Antonio Pescapé. Analysis of country-wide internet outages caused by censorship. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 1–18. ACM, 2011.

[89] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM, 2012.

[90] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318, 2013.

[91] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[92] Abhimanyu Das, Sreenivas Gollapudi, Emre Kıcıman, and Onur Varol. Information dissemination in heterogeneous-intent networks. In *Proceedings of the 8th ACM Conference on Web Science*, pages 259–268. ACM, 2016.

[93] Clayton A Davis, Giovanni Luca Ciampaglia, Luca Maria Aiello, Keychul Chung, Michael D Conover, Emilio Ferrara, Alessandro Flammini, Geoffrey C Fox, Xiaoming Gao, Bruno Gonçalves, et al. Osome: The iuni observatory on social media. Technical report, PeerJ Preprints, 2016.

[94] Clayton Allen Davis†, Onur Varol†, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[95] Richard Dawkins et al. *The selfish gene*. Oxford university press, 2016.

[96] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1431–1442. ACM, 2013.

[97] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM, 2013.

[98] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.

[99] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *ICWSM*, page 2, 2013.

[100] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[101] Kevin M DeLuca, Sean Lawson, and Ye Sun. Occupy wall street on the public screens of social media: The many framings of the birth of a protest movement. *Communication, Culture & Critique*, 5(4):483–509, 2012.

[102] Edwin Diamond and Stephen Bates. *The spot: The rise of political advertising on television*. Mit Press, 1992.

[103] Nicholas DiFonzo and Prashant Bordia. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35, 2007.

[104] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. *Available at SSRN: http://dx.doi.org/10.2139/ssrn.2235423*, 2013. Presented at 108th annual meeting of the American Sociological Association.

[105] Joan DiMicco, David R. Millen, Werner Geyer, Casey Dugan, Beth Brownholtz, and Michael Muller. Motivations for social networking at work. In *Proc. of the ACM conference on Computer supported cooperative work*, pages 711–720. ACM, 2008.

[106] Daniela V Dimitrova, Adam Shehata, Jesper Strömbäck, and Lars W Nord. The effects of digital media on political knowledge and participation in election campaigns: Evidence from panel data. *Communication Research*, 41(1):95–118, 2014.

[107] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, 2011.

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

[109] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, 11(2):415–441, 2006.

[110] Aviad Elyashar, Michael Fire, Dima Kagan, and Yuval Elovici. Homing socialbots: intrusion on a specific organization's employee using socialbots. In *Proc. IEEE/ACM Intl. Conf. on Advances in Social Networks Analysis and Mining*, pages 1358–1365, 2013.

[111] Shimon Even and Burkhard Monien. On the number of rounds necessary to disseminate information. In *Proceedings of the first annual ACM symposium on Parallel algorithms and architectures*, pages 318–327. ACM, 1989.

[112] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[113] Emilio Ferrara. Manipulation and abuse on social media. *SIGWEB Newsletter*, Spring(4):1–9, 2015.

[114] Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. *Proceedings of the 2013 IEEE/ACM International Conference on Advences in Social Network Analysis and Mining (ASONAM'13)*, 2013.

[115] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[116] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: social butterflies or frequent fliers? In *Proceedings of the first ACM conference on Online Social Networks (COSN)*, pages 213–222. ACM, 2013.

[117] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Detection of promoted social media campaigns. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[118] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. Predicting online extremism, content adopters, and interaction reciprocity. *arXiv preprint arXiv:1605.00659*, 2016.

[119] Emilio Ferrara and Zeyao Yang. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, 1:e26, 2015.

[120] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 745–754. ACM, 2011.

[121] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics, 2010.

[122] John C Fisher and Robert H Pry. A simple substitution model of technological change. *Technological forecasting and social change*, 3:75–88, 1971.

[123] James Montgomery Flagg. Uncle sam, 1917. [Online; accessed February 27, 2017].

[124] Office for Emergency Management. Office of War Information. Domestic Operations Branch. Bureau of Special Services. Stop this monster that stops at nothing. produce to the limit. this is your war, 1941. [Online; accessed February 27, 2017].

[125] Morgan R Frank, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Happiness and the patterns of life: A study of geolocated tweets. *Scientific reports*, 3, 2013.

[126] Kim Fridkin, Patrick J Kenney, and Amanda Wintersieck. Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising. *Political Communication*, 32(1):127–151, 2015.

[127] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *ICWSM*, 2014.

[128] R.K. Garrett. Protest in an information society: A review of literature on social movements and new ICTs. *Information, Communication & Society*, 9(02):202–224, 2006.

[129] Rumi Ghosh, Tawan Surachawala, and Kristina Lerman. Entropy-based classification of retweeting activity on twitter. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, August 2011.

[130] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *Proc. IEEE INFOCOM*, pages 1–9, 2010.

[131] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2015.

[132] Ted Goertzel. Belief in conspiracy theories. *Political Psychology*, pages 731–742, 1994.

[133] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE, 2011.

[134] Ann Goldberg. Reading and writing across the borders of dictatorship: Self-censorship and emigrant experience in nazi and stalinist europe. In *Letters across Borders*, pages 158–172. Springer, 2006.

[135] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[136] Nilüfer Göle. Gezi–anatomy of a public square movement. *Insight Turkey*, 15(3), 2013.

[137] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[138] Sandra González-Bailón, Javier Borge-Holthoefer, and Yamir Moreno. Broadcasters and hidden influentials in online protest diffusion. *American Behavioral Scientist*, 57(7):943–965, 2013.

[139] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific Reports*, 1, 2011.

[140] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. Assessing the bias in communication networks sampled from twitter. *arXiv preprint arXiv:1212.1684*, 2012.

[141] G Thomas Goodnight and John Poulakos. Conspiracy rhetoric: From pragmatism to fantasy in public discourse. *Western Journal of Speech Communication*, 45(4):299–316, 1981.

[142] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[143] Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. Extracting diurnal patterns of real world activity from social media. In *ICWSM*, 2013.

[144] Rosanna Guadagno and Robert Cialdini. Online persuasion and compliance: Social influence on the internet and beyond. *The social net: Human behavior in cyberspace*, pages 91–113, 2005.

[145] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[146] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[147] Trevor. Hastie, Robert. Tibshirani, and J Jerome H Friedman. *The elements of statistical learning*. Springer New York, 2001.

[148] Stefanie Haustein, Timothy D Bowman, Kim Holmberg, Andrew Tsou, Cassidy R Sugimoto, and Vincent Larivière. Tweets as impact indicators: Examining the implications of automated "bot" accounts on twitter. *Journal of the Association for Information Science and Technology*, 67(1):232–238, 2016.

[149] Alfred Hermida. From tv to twitter: How ambient news became ambient journalism. 2010.

[150] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[151] Bernie Hogan. The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, page 0270467610385893, 2010.

[152] Maureen Honey. *Creating Rosie the Riveter: class, gender, and propaganda during World War II.* Univ of Massachusetts Press, 1985.

[153] P.N. Howard, A. Duffy, D. Freelon, M. Hussain, W. Mari, and M. Mazaid. Opening closed regimes: What was the role of social media during the arab spring. Technical Report 2011.1, Project on Information Technology and Political Islam, 2011.

[154] Tim Hwang, Ian Pearce, and Max Nanis. Socialbots: Voices from the fronts. *Interactions*, 19(2):38–45, 2012.

[155] Roya Imani Giglou, Christine Ogan, and Leen d'Haenens. The ties that bind the diaspora to turkey and europe during the gezi protests. *New Media & Society*, 2017.

[156] Mohsen JafariAsbagh, Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media streams. *Social Network Analysis and Mining*, 4(1):1–13, 2014.

[157] Salman Jamali and Huzefa Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. In *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, pages 32–38. IEEE, 2009.

[158] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[159] J.C. Jenkins. Resource mobilization theory and the study of social movements. *Annual Review of Sociology*, 9:527–553, 1983.

[160] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 8. ACM, 2013.

[161] George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference*, pages 121–129, 1994.

[162] Adam N. Joinson. Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proc. of the SIGCHI conference on Human Factors in Computing Systems*, pages 1027–1036. ACM, 2008.

[163] Lynda Lee Kaid and Anne Johnston. Negative versus positive television advertising in us presidential campaigns, 1960–1988. *Journal of communication*, 41(3):53–064, 1991.

[164] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 667–677, 2013.

[165] Hyunjin Kang, Keunmin Bae, Shaoke Zhang, and S Shyam Sundar. Source cues in online news: Is the proximate source more powerful than distal sources? *Journalism & Mass Communication Quarterly*, 88(4):719–736, 2011.

[166] Anna Kata. Anti-vaccine activists, web 2.0, and the postmodern paradigm–an overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*, 30(25):3778–3789, 2012.

[167] Qing Ke. Sharing means renting?: An entire-marketplace analysis of airbnb. *arXiv preprint arXiv:1701.01645*, 2017.

[168] Qing Ke, Yong-Yeol Ahn, and Cassidy R Sugimoto. A systematic identification and analysis of scientists on twitter. *PloS one*, 12(4):e0175368, 2017.

[169] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

[170] Gary King, Jennifer Pan, and Margaret E Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107(02):326–343, 2013.

[171] Wolfgang H Kirchner. The evolution of communication. *Trends in Cognitive Sciences*, 1(9):353–353, 1997.

[172] Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, Catherine A Bliss, and Peter Sheridan Dodds. Positivity of the english language. *PLoS ONE*, 7(1):e29484, 2012.

[173] Jonas Krauss, Stefan Nann, Daniel Simon, Peter A Gloor, and Kai Fischbach. Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS*, pages 2026–2037, 2008.

[174] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P Gummadi. Geographic dissection of the Twitter network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[175] Mehmet Bariş Kuymulu. Reclaiming the right to the city: Reflections on the urban uprisings in turkey. *City*, 17(3):274–278, 2013.

[176] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *Proc. IEEE International Conference on Data Mining series (ICDM)*, 2013.

[177] Cliff Lampe, Nicole Ellison, and Charles Steinfield. A face (book) in the crowd: Social searching vs. social browsing. In *Proc. of the 20th anniversary conference on Computer supported cooperative work*, pages 167–170. ACM, 2006.

[178] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[179] T Lauricella, CHRISTOPHER S Stewart, and SHIRA Ovide. Twitter hoax sparks swift stock swoon. *The Wall Street Journal*, 23, 2013.

[180] Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *Proc. 5th AAAI Intl. Conf. on Web and Social Media*, 2011.

[181] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in Twitter. In *Proc. the 21th International Conference on World Wide Web*, pages 251–260, 2012.

[182] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 90–97, 2010.

[183] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506. ACM, 2009.

[184] Adrian Letchford, Helen Susannah Moat, and Tobias Preis. The advantage of short paper titles. *Royal Society Open Science*, 2(8):150266, 2015.

[185] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[186] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[187] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[188] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, 2003.

[189] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, 2007.

[190] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah boyd. The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5:1375–1405, 2011.

[191] Zongyang Ma, Aixin Sun, and Gao Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 64(7):1399–1410, 2013.

[192] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[193] Walid Magdy, Kareem Darwish, and Ingmar Weber. # failedrevolutions: Using twitter to study the antecedents of isis support. *arXiv preprint arXiv:1503.02401*, 2015.

[194] Darlene C Mahaney. Propaganda posters. *OAH Magazine of History*, 16(3):41–46, 2002.

[195] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the Twitter stream. In *Proceedings of the 2010 International Conference on Management of Data*, pages 1155–1158. ACM, 2010.

[196] Ian McAllister et al. The personalization of politics. *The Oxford handbook of political behavior*, pages 571–588, 2007.

[197] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908. International World Wide Web Conferences Steering Committee, 2013.

[198] J.D. McCarthy and M.N. Zald. Resource mobilization and social movements: A partial theory. *American Journal of Sociology*, 82(6):1212–1241, 1977.

[199] Anthony McCosker. Trolling as provocation youtube's agonistic publics. *Convergence: The International Journal of Research into New Media Technologies*, 20(2):201–217, 2014.

[200] David J McIver, Jared B Hawkins, Rumi Chunara, Arnaub K Chatterjee, Aman Bhandari, Timothy P Fitzgerald, Sachin H Jain, and John S Brownstein. Characterizing sleep issues using twitter. *Journal of medical Internet research*, 17(6), 2015.

[201] Yelena Mejova, Hamed Haddadi, Anastasios Noulas, and Ingmar Weber. # foodporn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 51–58. ACM, 2015.

[202] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we RT? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[203] Panagiotis Metaxas and Eni Mustafaraj. The rise and the fall of a citizen reporter. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 248–257. ACM, 2013.

[204] Panagiotis T Metaxas and Eni Mustafaraj. Social media and the elections. *Science*, 338(6106):472–473, 2012.

[205] Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. Using twittertrails. com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 69–72. ACM, 2015.

[206] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of Twitter users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[207] Lewis Mitchell, Kameron Decker Harris, Morgan R Frank, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417, 2013.

[208] Silvia Mitter, Claudia Wagner, and Markus Strohmaier. A categorization scheme for socialbot attacks in online social networks. In *Proc. of the 3rd ACM Web Science Conference*, 2013.

[209] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. The Twitter of Babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, January 2013.

[210] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[211] Merrill Morris and Christine Ogan. The internet as mass medium. *Journal of Computer-Mediated Communication*, 1(4):0–0, 1996.

[212] Chris Murray. *Popaganda: superhero comics and propaganda in World War Two.* Copenhagen: University of Copenhagen Press, 2000.

[213] Eni Mustafaraj and P Takis Metaxas. From obscurity to prominence in minutes: Political speech and real-time search. In *WebSci10: Extending the Frontiers of Society On-Line*, pages 317–323, 2010.

[214] D.J. Myers. Communication technology and social movements: Contributions of computer networks to activism. *Social Science Computer Review*, 12(2):250–260, 1994.

[215] Seth A Myers and Jure Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World wide web*, pages 913–924. ACM, 2014.

[216] Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012.

[217] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[218] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Yong-Yeol Ahn, and Alessandro Flammini. Information overload in group communication: From conversation to cacophony in the twitch chat. *arXiv preprint arXiv:1610.06497*, 2016.

[219] Azadeh Nematzadeh, Emilio Ferrara, Alessandro Flammini, and Yong-Yeol Ahn. Optimal network modularity for information diffusion. *Physical review letters*, 113(8):088701, 2014.

[220] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

[221] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media)*, 2011.

[222] Brendan Nyhan and Jason Reifler. The effect of fact-checking on elites: A field experiment on us state legislators. *American Journal of Political Science*, 59(3):628–640, 2015.

[223] Brendan Nyhan, Jason Reifler, and Peter A Ubel. The hazards of correcting myths about health care reform. *Medical care*, 51(2):127–132, 2013.

[224] Christine Ogan and Onur Varol. What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during gezi park. *Information, Communication & Society*, pages 1–19, 2016.

[225] Alexandra Olteanu, Onur Varol, and Emre Kıcıman. Towards an open-domain framework for distilling the outcomes of personal experiences from social media timelines. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[226] Alexandra Olteanu, Onur Varol, and Emre Kıcıman. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proc. of The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2017.

[227] J.P. Onnela, S. Arbesman, M.C. González, A.L. Barabási, and N.A. Christakis. Geographic constraints on social network groups. *PLoS ONE*, 6(4):e16939, 2011.

[228] Raj Kumar Pan, Kimmo Kaski, and Santo Fortunato. World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, 2, 2012.

[229] John Paparrizos, Ryen W White, and Eric Horvitz. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice*, page JOPR010504, 2016.

[230] M Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[231] Michael J Paul, Ryen W White, and Eric Horvitz. Search and breast cancer: On episodic shifts of attention over life histories of an illness. *ACM Transactions on the Web (TWEB)*, 10(2):13, 2016.

[232] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[233] Andrew Perrin. Social media usage. *Pew Research Center*, 2015.

[234] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pages 365–374. ACM, 2013.

[235] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. ACL, 2011.

[236] Daniele Quercia. Don't worry, be happy: The geography of happiness on facebook. In *Proceedings of ACM Web Science 2013*, 2013.

[237] Daniele Quercia, Licia Capra, and Jon Crowcroft. The social world of Twitter: Topics, geography, and emotions. In *Proc. of the 6th International AAAI Conference on Weblogs and Social Media*, pages 298–305, 2012.

[238] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE, 2011.

[239] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[240] J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proc. 5th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, pages 297–304, 2011.

[241] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proc. 20th Intl. ACM Conf. on World Wide Web (WWW)*, pages 249–252, 2011.

[242] Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*, 2014.

[243] Felice Resnik, Amy Bellmore, Jun-Ming Xu, and Xiaojin Zhu. Celebrities emerge as advocates in tweets about bullying. *Translational Issues in Psychological Science*, 2(3):323, 2016.

[244] Everett M Rogers and F Floyd Shoemaker. Communication of innovations; a cross-cultural approach. 1971.

[245] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–1123, 2008.

[246] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, 2009.

[247] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Jon Crowcroft. Track globally, deliver locally: improving content delivery networks by tracking geographic social

cascades. In *Proceedings of the 20th International Conference on World Wide Web*, pages 457–466. ACM, 2011.

[248] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. Distance matters: geo-social metrics for online social networks. *Proceedings of the 3rd Workshop on Online Social Networks*, 10, 2010.

[249] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media)*, pages 329–336, 2011.

[250] David Selassie, Brandon Heller, and Jeffrey Heer. Divided edge bundling for directional network data. *IEEE Trans. Visualization & Comp. Graphics*, 17:2354–2363, 2011.

[251] Pavel Senin and Sergey Malinchik. Sax-vsm: Interpretable time series classification using sax and vector space model. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1175–1180. IEEE, 2013.

[252] M.Á. Serrano, M. Boguñá, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

[253] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[254] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.

[255] Maeve Shearlaw. From britain to beijing: how governments manipulate the internet. Accessed online at http://www.theguardian.com/world/2015/apr/02/russia-troll-factory-kremlin-cyber-army-comparisons, April 2015.

[256] Michele J Shover. Roles and images of women in world war i propaganda. *Politics & Society*, 5(4):469–486, 1975.

[257] Thiago H Silva, Pedro OS de Melo, Jussara Almeida, Mirco Musolesi, and Antonio Loureiro. You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare. *arXiv preprint arXiv:1404.1009*, 2014.

[258] Craig Silverman. Lies, damn lies, and viral content. how news websites spread (and debunk) online rumors, unverified claims, and misinformation. *Tow Center for Digital Journalism*, 2015.

[259] Dennis M Simon and Charles W Ostrom. The impact of televised speeches and foreign travel on presidential approval. *Public Opinion Quarterly*, 53(1):58–82, 1989.

[260] Herbert A Simon. Designing organizations for an information-rich world. 1971.

[261] Brian Skyrms. *Signals: Evolution, learning, and information*. Oxford University Press, 2010.

[262] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[263] Didier Sornette, Fabrice Deschâtres, Thomas Gilbert, and Yann Ageon. Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Physical Review Letters*, 93(22):228701, 2004.

[264] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, volume 1, 2012.

[265] Pablo Suárez-Serrato, Margaret E Roberts, Clayton Davis, and Filippo Menczer. On the influence of social bots in online protests. In *International Conference on Social Informatics*, pages 269–278. Springer, 2016.

[266] V.S. Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, Rand Waltzman, Andrew Stevens, Alexander Dekhtyar, Shuyang Gao, Tad Hogg, Farshad Kooti, Yan Liu, Onur Varol, Prashant Shiralkar, Vinod Vydiswaran, Qiaozhu Mei, and Tim Huang. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.

[267] Jonathan Sullivan. A tale of two microblogs in china. *Media, Culture & Society*, 34(6):773–783, 2012.

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

[269] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*, 2014.

[270] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652. ACM, 2012.

[271] Z. Tufekci and C. Wilson. Social Media and the Decision to Participate in Political Protest: Observation from Tahrir Square. *Journal of Communication*, 62:363–379, 2012.

[272] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proc. ICWSM*, 10:178–185, 2010.

[273] Twitter. Twitter transparency reports. `https://transparency.twitter.com/removal-requests`. Accessed: 2016-01-10.

[274] Twitter Inc. Faqs about trends on twitter. Accessed online at https://support.twitter.com/articles/101125, July 2016.

[275] Unknwon. Together we can do it! - keep 'em firing, 1941. [Online; accessed February 27, 2017].

[276] Umut Uras. What inspires turkey's protest movement. *Al Jazeera*, June 5, 2013.

[277] U.S. Securities and Exchange Commission. Updated investor alert: Social media and investing — stock rumors. Accessed online at http://www.sec.gov/oiea/investor-alerts-bulletins/ia_rumors.html, November 2015.

[278] S. Valenzuela, A. Arrigada, and A. Scherman. The Social Media Basis of Youth Protest Behavior: The Case of Chile. *Journal of Communication*, 62(2):299–314, 2012.

[279] J. Van Laer and P. Van Aelst. Cyber-protest and civil society: the internet and action repertoires in social movements. *Handbook on Internet Crime*, pages 230–254, 2009.

[280] Onur Varol. Raw eeg data classification and applications using svm. *İstanbul Technical University*, 2010.

[281] Onur Varol. Spatiotemporal analysis of censored content on twitter. In *Proceedings of the 8th ACM Conference on Web Science*, pages 372–373. ACM, 2016.

[282] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*, 2017.

[283] Onur Varol, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. Early detection of promoted campaigns on social media. *arXiv preprint arXiv:1703.07518*, 2017.

[284] Onur Varol, Emilio Ferrara, Christine L Ogan, Filippo Menczer, and Alessandro Flammini. Evolution of online user behavior during a social upheaval. In *Proceedings of the 2014 ACM conference on Web science*, pages 81–90. ACM, 2014.

[285] Onur Varol and Filippo Menczer. Connecting dream networks across cultures. In *Proceedings of the companion publication of the 23rd international conference on World Wide Web companion*, pages 1267–1272. International World Wide Web Conferences Steering Committee, 2014.

[286] Onur Varol, Deniz Yuret, Burak Erman, and Alkan Kabakçıoğlu. Mode coupling points to functionally important residues in myosin ii. *Proteins: Structure, Function, and Bioinformatics*, 82(9):1777–1786, 2014.

[287] John-Paul Verkamp and Minaxi Gupta. Inferring mechanics of web censorship around the world. In *FOCI*, 2012.

[288] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425, 2009.

[289] Juha Antero Vuori and Lauri Paltemaa. The lexicon of fear: Chinese internet control practice in sina weibo microblog censorship. *Surveillance & Society*, 13(3/4):400–421, 2015.

[290] Claudia Wagner, Silvia Mitter, Christian Körner, and Markus Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. In *Proceedings of the 21th International Conference on World Wide Web*, pages 41–48, 2012.

[291] Randall Wald, Taghi M Khoshgoftaar, Antonio Napolitano, and Chris Sumner. Predicting susceptibility to social bots on twitter. In *Proc. 14th Intl. IEEE Conf. on Information Reuse and Integration (IRI)*, pages 6–13, 2013.

[292] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y Zhao. You are how you click: Clickstream analysis for sybil detection. In *Proc. USENIX Security*, pages 1–15. Citeseer, 2013.

[293] Gang Wang, Manish Mohanlal, Christo Wilson, Xiao Wang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. Social turing tests: Crowdsourcing sybil detection. In *Proc. of the 20th Network & Distributed System Security Symposium (NDSS)*, 2013.

[294] Senzhang Wang, Zhao Yan, Xia Hu, Philip S Yu, and Zhoujun Li. Burst time prediction in cascades. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 325–331. AAAI Press, 2015.

[295] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, pages 1–17, 2013.

[296] L Weng, A Flammini, A Vespignani, and F Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2, 2012.

[297] L Weng, F Menczer, and YY Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3:2522–2522, 2012.

[298] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3, 2013.

[299] Lilian Weng, Jacob Ratkiewicz, Nicola Perra, Bruno Gonçalves, Carlos Castillo, Francesco Bonchi, Rossano Schifanella, Filippo Menczer, and Alessandro Flammini. The role of information diffusion in the evolution of social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 356–364. ACM, 2013.

[300] Darrell M West. *Air Wars: Television Advertising and Social Media in Election Campaigns, 1952-2016*. CQ Press, 2017.

[301] Ryen W White, Sheng Wang, Apurv Pant, Rave Harpaz, Pushpraj Shukla, Walter Sun, William DuMouchel, and Eric Horvitz. Early identification of adverse drug reactions from search log data. *Journal of biomedical informatics*, 59:42–48, 2016.

[302] Wikipedia. Cities and metropolitan areas of the United States. `http://en.wikipedia.org/wiki/Cities_and_metropolitan_areas_of_the_United_States`, 2012.

[303] Wikipedia. List of the busiest airports in the United States. `http://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States`, 2012.

[304] R HAVEN Wiley. The evolution of communication: information and manipulation. *Animal behaviour*, 2:156–189, 1983.

[305] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. ACL Conf. on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

[306] Wendy Wood. Attitude change: Persuasion and social influence. *Annual review of psychology*, 51(1):539–570, 2000.

[307] S. Wray. On electronic civil disobedience. *Peace Review*, 11(1):107–111, 1999.

[308] Fang Wu and Bernardo A Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104:17599–17601, 2007.

[309] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7):589–600, 2014.

[310] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[311] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y Zhao, and Yafei Dai. Uncovering social network sybils in the wild. *ACM Trans. Knowledge Discovery from Data*, 8(1):2, 2014.

[312] Katherine K Young and Paul Nathanson. *Sanctifying misandry: Goddess ideology and the fall of man*. McGill-Queen's Press-MQUP, 2010.

[313] Lena L Zhang. Behind the 'great firewall'decoding china's internet media policies from the inside. *Convergence: The International Journal of Research into New Media Technologies*, 12(3):271–291, 2006.

[314] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International*

Conference on World Wide Web, pages 1395–1405. International World Wide Web Conferences Steering Committee, 2015.

[315] Tao Zhu, David Phipps, Adam Pridgen, Jedidiah R Crandall, and Dan S Wallach. Tracking and quantifying censorship on a chinese microblogging site. *arXiv preprint arXiv:1211.6166*, 2012.

[316] Tao Zhu, David Phipps, Adam Pridgen, Jedidiah R Crandall, and Dan S Wallach. The velocity of censorship: High-fidelity detection of microblog post deletions. *arXiv preprint arXiv:1303.0597*, 2013.

Onur VAROL
www.onurvarol.com School of Informatics and Computing Indiana University
ovarol@indiana.edu 919 E. 10th St, Bloomington IN, 47408, USA

Research Interests

Complex Systems, Network Science, Data Science, Machine Learning, Computational Social Science

Education

Indiana University, Bloomington, Indiana, USA
Ph.D., Complex Systems track of the Ph.D. in Informatics and minor in Statistics, (June 2017)
Dissertation: Analyzing Social Big Data to Study Online Discourse and its Manipulation
Committee: Filippo Menczer (chair), Alessandro Flammini, Yong-Yeol Ahn, Christine L. Ogan, Weihua An

Koç University, Istanbul, Turkey
M.Sc., Computer Science and Engineering, (July 2012)
Thesis: Modal analysis of Myosin II and Identification of Functionally Important Sites
Advisors: Deniz Yüret, Alkan Kabakçıoglu

Istanbul Technical University, Istanbul, Turkey
B.Sc., Electronics Engineering, (June 2010)
B.Sc., Physics Engineering, (June 2012)

Publications

Journal Articles

J.8 **O Varol**. **"Deception Strategies and Threats for Online Discussions"**. (under review)

J.7 **O Varol**, E Ferrara, F Menczer, A Flammini. **"Early Detection of Promoted Campaigns on Social Media"**. (under review)

J.6 C. Ogan, **O. Varol "Combining Content Analysis with Network Analysis in the Study of a Social Movement: Twitter Use During Gezi Park Protests"**, *Information, Communication and Society* 1-19, 2016

J.5 E Ferrara, **O Varol**, C Davis, F Menczer, A Flammini **"The Rise of Social Bots"**, *Communications of the ACM* 59(7):96-104, 2016

J.4 Davis CA, Ciampaglia GL, Aiello LM, Chung K, Conover MD, Ferrara E, Flammini A, Fox GC, Gao X, Gonçalves B, Grabowicz PA, Hong K, Hui P, McCaulay S, McKelvey K, Meiss MR, Patil S, Peli Kankanamalage C, Pentchev V, Qiu J, Ratkiewicz J, Rudnick A, Serrette B, Shiralkar P, **Varol O**, Weng L, Wu T, Younge AJ, Menczer F. **"OSoMe: The IUNI observatory on social media"**, *PeerJ Computer Science* 2: e87, 2016

J.3 V.S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, F. Menczer, R. Waltzman, A. Stevens, A.

Dekhtyar, S. Gao, T. Hogg, F. Kooti, Y. Liu, **O. Varol**, P. Shiralkar, V. Vydiswaran, Q. Mei, T. Huang. **"The DARPA Twitter Bot Challenge"**, *IEEE Computer* 49(6), 2016

J.2 M JafariAsbagh, E Ferrara, **O Varol**, F Menczer, A Flammini **"Clustering memes in social media streams"**, *Social Network Analysis and Mining* 4(237):1-13, 2014

J.1 **O. Varol**, D. Yuret, B. Erman, A. Kabakçıoglu **"Mode coupling points to functionally important residues in Myosin II"**, *PROTEINS: Structure, Function, and Bioinformatics* 82(9):1777-86, 2014

Conference Proceedings

C.13 **Varol O.**, Davis C., Ferrara E., Menczer F., Flammini A. **"Online Human-Bot Interactions: Detection, Estimation, and Characterization"**, ICWSM'17

C.12 Olteanu, A., **Varol, O.**, Kiciman, E. **"What Does Social Media Say about the Outcomes of Personal Experiences"**, CSCW'17

C.11 Ferrara E., Wang W., **Varol O.**, Flammini A., Galstyan A. **"Predicting online extremism, content adopters, and interaction reciprocity"**, SocInfo'16

C.10 **O Varol**, **"Spatiotemporal Analysis of Censored Content on Twitter"**. WebScience'16

C.9 A Das, S Gollapudi, E Kiciman, **O Varol**, **"Information Dissemination in Heterogeneous-Intent Networks"**. WebScience'16

C.8 E. Ferrara, **O Varol**, F Menczer, and A Flammini. **"Detection of Promoted Social Media Campaigns"**. ICWSM'16

C.7 A Olteanu, **O Varol**, E Kiciman. **"Towards an Open-Domain Framework for Distilling the Outcomes of Personal Experiences from Social Media Timelines"**. ICWSM'16

C.6 C Davis†, **O Varol**†, E Ferrara, A Flammini, F Menczer **"BotOrNot: A System to Evaluate Social Bots"**. WWW'16 Developers Day

C.5 **O Varol**, E Ferrara, C Ogan, F Menczer, and A Flammini. **"Evolution of online user behavior during a social upheaval"**. ACM Web Science Conference 2014 (**Best paper award**)

C.4 **O Varol** and F Menczer. **"Connecting Dream Networks Across Cultures"**. WWW 2014 workshop on "Connecting Online & Offline Life" (COOL)

C.3 Ferrara, E., **Varol, O.**, Menczer, F. & Flammini, A. **"Traveling Trends: Social Butterflies or Frequent Fliers?"**, ACM Conference on Online Social Networks (COSN 2013)

C.2 E Ferrara, M JafariAsbagh, **O Varol**, V Qazvinian, F Menczer, A Flammini **"Clustering Memes in Social Media"**, ASONAM 2013

C.1 Yasemin Alban; Tuba Ayhan; **Onur Varol**; Müştak Erhan Yalçın **"A Feature Filtering Method for EEG Data Classification"**, Signal Processing and Communications Applications (SIU), IEEE 19th Conference, Antalya, April 20-22 2011.

Conducted Research and Academic Experience

Research

DOISAC: Project name stands for "Detecting Orchestrated Information and Synthetic Account Campaigns". This project founded by the Office of Naval Research aims at detecting orchestrated information and synthetic activity campaigns on social media using machine learning and computational tools. In this project, I studied individual and group activities of terrorist recruiters. We build predictive models to identify accounts with malicious intentions and activities.

PIs: Dr. Emilio Ferrara, Assoc. Prof. Alessandro Flammini,

DESPIC: Project name stands for "Detecting Early Signature of Persuasion in Information Cascades" and aims to design a system detect persuasion campaigns at their early stage of inception, in the context of online social media. Our team built a system that analyzes social media data and extracts network, temporal, content, and user-based features to detect online campaigns. I worked on several modules of this framework: (i) a clustering procedure that uses metadata to compute similarity between memes; (ii) a classification system that determines whether a meme is potentially an orchestrated campaign or a genuine, grassroots conversation; (iii) a social bot detection framework called BotOrNot. This project founded by DARPA SMISC program.

PIs: Assoc. Prof. Alessandro Flammini, Prof. Fil Menczer

Truthy: This project aims to understand how information propagates through complex socio-technical information networks. In this project, I analyzed and studied roles of individuals during social upheavals, diffusion of trending topics in spatio-temporal space, and characterization of social media censorship and its effect on user behavior.

PIs: Assoc. Prof. Alessandro Flammini, Prof. Fil Menczer

Research on Modal analysis of Myosin II and Identification of Functionally Important Sites: During my Master's studies in Koc University, I worked on analysis of protein fluctuations to identify functionally important residues as my thesis project.

Advisors: Assoc. Prof. Deniz Yüret, Assist. Prof. Alkan Kabakçıoglu

Research on Modeling of Social Networks and Phase Transitions of Complex Systems (Graduation project for B.Sc in Physics Engineering)

Advisor: Prof. Ayse Erzan

Research on EEG Signal Processing and Classification (Graduation project for B.Sc in Electronics Engineering)

Advisor: Prof. Müstak Erhan Yalçın

Teaching Assistant

Indiana University, Bloomington IN, USA

Topics in Informatics: Performance Analytics (Spring 2017)

Koç University, Istanbul, Turkey

Machine Learning (Spring 2012)
Microprocessors (Fall 2011)
Probability and Random Variables (Spring 2011)
Discrete Mathematics (Fall 2010)

Work Experience

Microsoft Research, Redmond WA, USA

Research Intern (June 2015 - September 2015): I worked in CLUES group at MSR. I studied social media timelines of individuals to detect experiential activities and analyzed outcomes of those actions. This projects involves analyzing search query logs and social media timelines.

Microsoft Research, Redmond WA, USA

Research Intern (June 2014 - September 2014): I worked in CLUES group at MSR. I studied how individuals in social networks adopts their behavior to match with their inner intents. I carried out experiments on Amazon Mechanical Turk platform to justify our hypothesis in micro level and analyzed Twitter data to validate effects on macro level.

Stonefish Software Consultant, Istanbul, Turkey

Software Developer (August 2009 - March 2010): I worked on an enterprise web application development using C♯, ASP.NET, MSSQL.

Honors and Awards

Best paper award (WebSci'14)
Best poster award (CCS'15)
Travel grants: SIGWEB for WebSci'16 (750$), ICWSM'16 (350$), ACM for COSN'14 (1,500$), IU RKCSI for INFORMS'16 (500$)
Research Assistantship Indiana University (2012-2016)
Scholarship from TUBITAK (2010-2012)
Graduate scholarship from Koç University (2010-2012)
Electrical Engineering Society (EMO) Project Competition 1^{st} place (2010)
Bosch Scholarship for Undergraduate Education (2009)
Istanbul Technical University Student Council Vice President (2009)

Talks and Events

Invited Talks

- The Impact of Censorship on Tweeting Behaviors, Indiana University Conference on Big Data and Network Science (03/23/2017)
- Studying Individuals and Groups using Online Data, Northeastern University Network Science Institute (03/09/2017)
- Analysis of Online Discourse and Information Diffusion, INFORMS Meeting Nashville (11/15/2016)
- Detection of Online Manipulation, UND Big Data Summit Event (04/07/2016)
- Twitter Applications: Industry Panel Speaker, UND Big Data Summit Event (04/07/2016)

- Campaign and Social Bots Detection on Social Networks, Workshop in Network Science (WINS) Indiana University (02/18/2016)
- Evolution of online user behavior during a social upheaval, Indiana University Turkish Flagship Center (01/21/2015)
- Studying Social Dynamics Through Social Media, Istanbul Technical University Physics Department Colloquia (04/28/2014)

Computer and Language Skills

Programming Languages and Skills:

Frequent user of Python for data analysis using Matplotlib, NetworkX, Pandas, Scikit-learn, etc.
Experience in LaTeX, C / C++, C#, Java, MATLAB, R, OpenBUGS
Familiar with HTML, CSS, JS for frontend, Flask, Django, and ASP.NET for larger applications
Used MySQL, NoSQL (CouchDB and Riak), Map-Reduce
View projects on Github: github.com/onurvarol

Languages:

English (fluent), German (beginner), Turkish (native)

Community Service

Workshop co-chair: Open Science for an Open Society (CCS'16)
Conference PC member: CSCW'18, SocInfo'17, IC2S2'17, HyperText'17, NetSci'17, WWW'17, ICWSM'17
Conference subreviewer: ICWSM'15, IC2S2'15, ASONAM'15, WebScience'14
Journal reviews: PLoS One, Network Science, IEEE TKDE

Interests

I enjoy playing basketball and foosball. I also love traveling a lot and feeling wanderlust, you can visit my online travel journal (onurvarol.com/my_travels).