Online Social Network Analysis: A Survey of Research Applications in Computer Science

Article · April 2015

CITATIONS
9

READS
284

3 authors:

David Burth Kurka
Imperial College London
5 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Alan Godoy
CPqD
19 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Fernando J. Von Zuben
University of Campinas
343 PUBLICATIONS **6,144** CITATIONS
SEE PROFILE

Online Social Network Analysis: A Survey of Research Applications in Computer Science

David Burth Kurka DKURKA@DCA.FEE.UNICAMP.BR
Laboratory of Bioinformatics and Bioinspired Computing
University of Campinas
Campinas, São Paulo, Brazil

Alan Godoy GODOY@DCA.FEE.UNICAMP.BR
CPqD Foundation
Campinas, São Paulo, Brazil
and
Laboratory of Bioinformatics and Bioinspired Computing
University of Campinas
Campinas, São Paulo, Brazil

Fernando J. Von Zuben VONZUBEN@DCA.FEE.UNICAMP.BR
Laboratory of Bioinformatics and Bioinspired Computing
University of Campinas
Campinas, São Paulo, Brazil

Abstract

The emergence and popularization of online social networks suddenly made available a large amount of data from social organization, interaction and human behaviour. All this information opens new perspectives and challenges to the study of social systems, being of interest to many fields. Although most online social networks are recent (less than fifteen years old), a vast amount of scientific papers was already published on this topic, dealing with a broad range of analytical methods and applications. This work describes how computational researches have approached this subject and the methods used to analyse such systems. Founded on a wide though non-exaustive review of the literature, a taxonomy is proposed to classify and describe different categories of research. Each research category is described and the main works, discoveries and perspectives are highlighted.

Keywords: Online Social Networks, Survey, Computational Research, Machine Learning, Complex Systems

1. Introduction

One of the most revolutionary aspects of the Internet is, beyond the possibility of connecting computers from the entire world, the power to connect people and cultures. More and more the Internet is used for the development of online social networks (OSNs)—an adaptation of social organizations to the "virtual world". Currently, OSNs such as *Twitter*[1], *Google+*[2] and *Facebook*[3] have hundreds of millions of users (Ajmera, 2014). Futhermore, the average browsing time inside those services is increasing (Benevenuto et al., 2009) and many websites are featuring some sort of integration with social networking services. Although the effects of such services on personal interactions, cultural and living standards, education and politics are visible, understanding the whole extent of the influence and impact of those services is a challenging task.

The study of social networks is not something new. Since the emergence of the first human societies, social networks have been there forging individual and collective behavior. In the academia, research on social networks can be traced to the first decades of the twentieth century (Rice, 1927), while probably the most influential early work on social network analysis was the seminal paper "Contacts and Influence" (de Sola Pool and Kochen, 1978), written in the 1950's[4].

In recent years, however, with the popularization of OSNs, this research subject gained new momentum as new possibilities of study have arisen and plenty of data on social relations and interactions have become available. Even though the most popular OSNs have barely ten years of existence—Facebook was founded in 2004, Twitter in 2006 and Myspace[5] in 2003—, the volume of scientific work having them as subject is considerable. Finding order and sense among all the work produced is becoming a huge task, specially for new researchers, as the amount of produced material accumulates.

With this in mind, this work aims to present an introductory overview of research in online social network analysis, mapping the main areas of research and their perspectives. A comprehensive approach is taken, prioritizing the diversity of applications, but endeavouring to select relevant work and to analyse their actual contributions. Also, although many disciplines have been interested in this topic—it is possible to find related works in psychology, sociology, politics, economics, biology, philosophy, to name a few—, the present work will focus predominantly in computational approaches.

This work is structured as follows: in section 2 the main reasons and attractives of OSN research are discussed; in section 3 a proposal for a taxonomy is presented and sections 4, 5 and 6, following the proposed nomenclature, detail the main references and findings for each topic. Finally, in section 7 we conclude by presenting general remarks regarding the current stage of the research and a brief analysis of future perspectives.

1. `https://twitter.com`
2. `https://plus.google.com`
3. `https://www.facebook.com`
4. Despite being formally published only in 1978, early versions of this paper circulated among scholars since it was written. These early versions had strong impact on many researchers, including Stanley Milgram in his paper about the small-world phenomenon.
5. `https://myspace.com`

2. Online social networks as object of study

In this section, we make a brief introduction to the research about online social networks, discussing the reasons why this area is getting a very strong momentum, the kind of data being explored in the field and the computational tools commonly used by researchers to analyze social networks data.

2.1 Why should anyone research OSNs?

The attention given by the media and general public to OSNs can be a good motivation to justify the research in this field. However, from a computational perspective, OSNs present some particularities that must be taken into account, in order to understand researchers interests. The main reasons are listed below:

Data availability: every day, a huge amount of information travels through OSNs and much of it is freely available for researchers[6]. The current abundance of data has no precedent in the study of social systems and serves as basis for computational analysis and scientific work. Due to its large scale, social data can fit in the context of *big data* research.

Multiple authorship: differently from other corpora, the textual content produced in OSNs have different authorial sources. This enhances the information content and diversity of the data collected, presenting various styles, forms, contexts and expression strategies. Thereby, OSNs can be a rich repository of text for natural language processing applications.

Agent interaction: every individual user that composes such networks is an agent able to take decisions and interact with other users. This complex interaction dynamics produces effects that puzzle and interest several researchers.

Temporal dynamics: the fact that social data is generated continuously along time, allows analysis that take into account spatio-temporal processes and transformations, such as topic evolution or collective mobilization.

Instantaneity: besides the continuous generation, the social data is also provided at every moment, instantaneously. Thus, OSNs typically react in real time to both internal and external stimuli.

Ubiquity: following the technological development, which increases people's access to means of communication and information (as smartphones, tablets), OSNs content can be generated, virtually, anywhere and at any time. Also, data's *geolocation*, a feature present in many OSNs, add new possibilities to the analysis.

2.2 Which networks are explored?

Two main characteristics can be taken into consideration, before choosing a network to study: popularity (number of active users) and how easy is the data access.

6. Respecting, however, specified privacy limits and download rates.

Currently, the largest online social network is *Facebook*, with over one billion active users (Facebook, 2014). Although the use of data extracted from Facebook is present in literature (Dow and Friggeri, 2013; Kumar, 2012; Sun et al., 2009), the high proportion of protected content—generally due to users' privacy settings—severely restricts the analysis using this OSN as source.

Twitter, a popular microblogging tool (Cheong and Ray, 2011), can be considered by far the most studied OSN (Rogers, 2013). The existence of a well-defined public interface for software developers[7] to extract data from the network, the simplicity of its protocol[8] and the public nature of most of its content can be a good explanation for that. However, since the beginning of the service, rate policies have been created to control the amount of data allowed to be collected by researchers and analysts. This had a direct impact on research, as initial works had access to all the content published in the network, while today's works are usually limited by those policies (Rogers, 2013).

It is also worth mentioning the existence of Chinese counterpart services for Facebook and Twitter, like Sina-Weibo[9], the largest one, with more than 500 million registered users (Ong, 2013). Although the usage of those services may differ due to cultural aspects (Asur et al., 2011; Gao et al., 2012), similar lines of inquires can be developed in both the western and eastern equivalents (e.g.: Guo et al., 2011; Qu et al., 2011; Yang et al., 2012; Bao et al., 2013).

Other web services that integrate social networking features have been the focus of studies. Examples are media sites like YouTube[10] (Mislove et al., 2007) and Flickr[11] (Cha et al., 2009; Kumar et al., 2010b), and news services as Digg[12] (Hogg and Lerman, 2009a; Wu and Huberman, 2007). Research was also made with implicit social networks as email users (Tyler et al., 2005), university pages (Adamic and Adar, 2003, 2005) or blogs (Gruhl et al., 2004), even before the creation of social networking services.

2.3 Computational tools

There are, currently, many computational tools that help in the task of analyzing large social networks, like graph-based databases (e.g.: AllegroGraph[13] and Neo4J[14]), libraries to access online social networks APIs (e.g.: Instagram Ruby Gem[15] and Tweepy[16]), graph drawing softwares (e.g.: Graphviz[17] and Tulip[18]) and tools for graph manipulation and statistical analysis of networks. The present section, however, will focus only in this last

7. `https://dev.twitter.com`
8. In Twitter, users can post only 140 characters text messages, unlike Facebook, where users can send photos, videos and large text messages.
9. `http://weibo.com`
10. `https://www.youtube.com`
11. `https://www.flickr.com`
12. `http://digg.com`
13. `http://franz.com/agraph/allegrograph/`
14. `http://neo4j.com/`
15. Instagram Ruby Gem is an official Ruby wrapper for Instagram APIs, available at `https://github.com/Instagram/python-instagram`.
16. Tweepy is a third-party Python library to access Twitter API. Available at `http://www.tweepy.org/`.
17. `http://www.graphviz.org/`
18. `http://tulip.labri.fr/`

category, as it is more relevant to the kind of analysis conduced in the studies presented in this survey.

Even when considering only tools for graph analysis and manipulation, there are dozens of alternatives, ranging from general purpose graph libraries to advanced commercial tools aimed at specific business. For an extensive list of social networks analysis software, we refer to Wikipedia's entry on the subject (Wikipedia, 2015).

When considering applications commonly used in academic works, a division in two groups of tools is clear: (a) graphical user interface (GUI), which are based stand-alone software, focusing on ease of use by non-programmers, and (b) programming language libraries, that are usually more flexible and have more functionalities.

In the first group, the most widely adopted tool is Gephi[19] (Bastian et al., 2009), which is a Java-based open source software licensed under the Common Development and Distribution License (CDDL) and GNU General Public License (GPL). Gephi is able to deal with large graphs (up to 1 million nodes and edges, according to its website), allowing node/edge filtering. It features diverse algorithms to draw graphs, detect communities, generate random graphs and calculate network metrics, like centrality measures (e.g.: betweenness, closeness and PageRank), diameter, and clustering coefficient. It is also able to deal with temporal information and hierarchical graphs and has support for third-party plugins. In addition to the stand-alone software, Gephi is also available as a Java module through Gephi Toolkit[20].

Another GUI-based software worth mentioning is Cytoscape[21] (Saito et al., 2012), also open source and licensed under the GNU Lesser General Public License (LGPL). As Gephi, Cytoscape is written in Java and offers graph drawing, community detection algorithms, network metrics, node/edge filtering and it also supports plugins. Despite being intended for the analysis of biomolecular networks, Cytoscape can be used to analyze graphs from any kind of source, including social networks.

The most adopted and feature-rich libraries in the second group are NetworkX and igraph. Both libraries can handle millions of nodes and edges (Akhtar et al., 2013) and offer advanced algorithms for networks, as checking isomorphisms, searching for connected components, cliques, communities and k-cores, and calculating dominating and independent sets and minimum spanning trees.

NetworkX[22] (Hagberg et al., 2008) is an open source project—under the Berkeley Software Distribution license (BSD)—sponsored by Los Alamos National Lab, which is in active development since 2002. Despite the recurrent addition of new functionalities, it is a very stable library, as it includes extensive unit-testing. NetworkX is fully implemented in Python and is interoperable with NumPy and SciPy, the language's standard packages for advanced mathematics and scientific computation. It also has remarkable flexibility: nodes can be almost anything—texts, numbers, images and even other graphs—and graphs, nodes and edges can have attributes of any type. The library can deal not only with common graphs, but also with digraphs, multigraphs and dynamic graphs. Among the specific features of

19. https://gephi.github.io/
20. http://gephi.github.io/toolkit/
21. http://www.cytoscape.org/
22. https://networkx.github.io/

NetworkX are a particularly large set of graph generators and a number of special functions for bipartite graphs.

igraph[23] (Csardi and Nepusz, 2006) is a performance-oriented graph library written in C with official interfaces for C, Python and R and a third-party binding for Ruby. If on the one hand it is not as flexible as NetworkX, on the other hand it can be even 10 times faster when performing some functions (Akhtar et al., 2013). Many advanced network analysis methods are available in igraph, including classical techniques from sociometry, like dyad and triad census and structural holes scores, and more recent methods, like motif estimation, decomposing a network into graphlets and different algorithms for community detection. As all other tools presented in this section, igraph is an open source project (it is licensed under the GNU GPL).

Two more libraries worth citing are graph-tool[24] and NetworKit[25], open source frameworks intended to be much faster than mainstream alternatives by making intensive use of parallelism. Both libraries are implemented mostly in C++ and have Python APIs providing broad lists of functionalities, though not as comprehensive as NetworkX and igraph's. graph-tool (Peixoto, 2014) is licensed under the GNU GPL and is developed since 2006. NetworKit (Staudt et al., 2014) is very recent: it was created in 2013 in the Karlsruhe Institute of Technology, in Germany. It is under the MIT license and is designed to be interoperable with NetworkX. Differently from other libraries, it aims at networks with billions of nodes and edges and is particularly well-suited for high-performance computing.

The libraries discussed here implement a vast range of graph functions. Some of these functions, however, are not available in all tools. We recommend that researchers in need of specific functionalities to check the libraries' documentation, available at their websites. All these libraries are under active development and are well documented. For more complete comparisons between network libraries, we refer to Combe et al. (2010); Akhtar et al. (2013); Staudt et al. (2014).

3. Categories of study

In order to simplify the presentation of the wide range of works devoted to the analysis of Online Social Networks, a categorization of the areas of research is needed. Here we will propose a taxonomy that covers different aspects of this research, structuring all the surveyed works in three main groups: (a) structural analysis, (b) social data analysis and (c) social interaction analysis. Fig. 1 illustrates this structure, with its respective subdivisions.

Structural analysis is the earliest category of study, since it contemplates initial inquiries about the structure and functionality of social networking services (SNSs), as they were launched. Researchers were interested in simply knowing what are those services and why so many people were being attracted to them. Also, the huge structures that were being formed proved to be worthy investigating and comparing to other known networks (as biological and offline social networks). Despite a decrease in the number of works in recent years, this area of research is still very active.

23. `http://igraph.org/`
24. `http://graph-tool.skewed.de/`
25. `http://networkit.iti.kit.edu/`

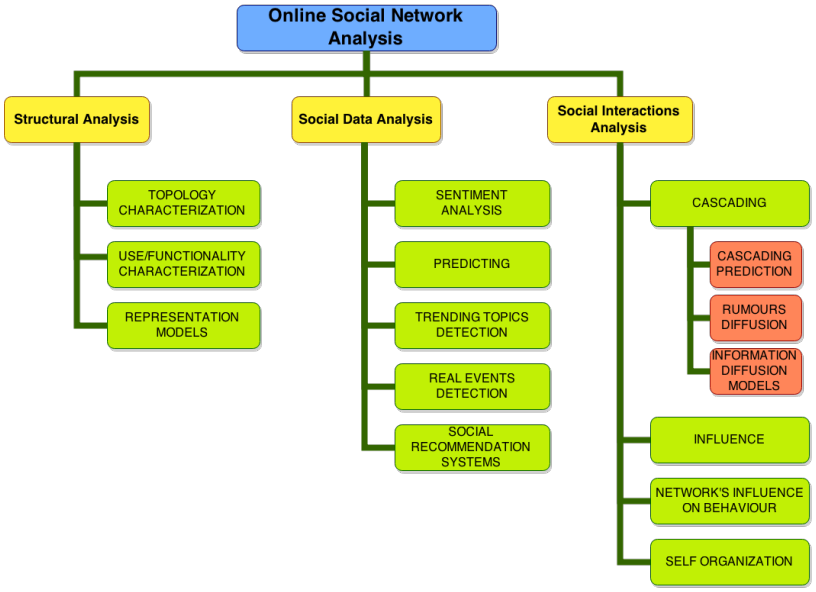

Figure 1: Categories of study on Online Social Networks, from a computational perspective.

Social data analysis represents a second branch, in which researchers, more used to what OSNs are, started to use and analyse what OSNs *produce*. This area, dominated by data scientists, exploits the huge amount of rich data produced by OSNs to do all kinds of applications. Usually only the data produced by users is considered, not having much importance the topology of users' connections or other network features.

Finally, *social interaction analysis* can be considered the most challenging and uncertain area, as it deals with the individuals using the SNSs. Using all the rich data provided by OSNs, such as users' friendships and the record of social interaction, it is possible to have insights on human behaviour. By analysing and interpreting this data, computer scientists are able to make discoveries and also collaborate with other fields of research, such as psychology, sociology and even biology.

We are unaware of other works that propose a taxonomy for the computational study of OSNs in general. However, previous works were made specifically focusing on studies about Twitter. Memon and Alhajj (2010) and Cheong and Ray (2011) tracked papers produced from 2008 to 2010 and found categories very similar to the ones presented above. However, their general classification is based on only two main areas: user domain and message domain. Williams et al. (2013) systematically collected all the research papers since 2011 containing the word "Twitter", and defined four main aspects: message, user, technology and concept. Message could be related to social data analysis, user to social interaction analysis and technology and concept to structural analysis. However, that work did not further deepen the classification in subcategories.

Another interesting perspective is the study conducted by Rogers (2013), which described the evolution of Twitter and how it has been attracting researchers. According to him, Twitter passed through three phases: Twitter I, when the service was used mainly

to connect people, but dominated by banality; Twitter II, a more mature network, able to promote and organize mobilizations; and Twitter III, a historical valuable big database used to understand society and the recent past.

Of course, we do not expect to achieve consensus with this taxonomy. Imposing categories to any study can be helpful for contextualization, but can also be misleading and endowed with some degree of arbitrariness. Also, works can belong to more than one category and there can be some intersection between different areas of research. The aim of this survey, therefore, is to serve as an introductory overview of the current status of the field, supported by the proposed taxonomy.

4. Structural analysis

Under structural analysis are works that have OSNs structure and operation as objects of study. Many can be the reasons researchers are interested in the study of a network: to understand how it is composed, to compare its structure to other known networks (specially with offline social networks) or to create models of social organization.

Since the end of the last century, studies showed that many real networks have some non-trivial properties, such as small average distances between nodes[26] (Watts and Strogatz, 1998) and number of connections per node following a power-law[27] (Barabási, 1999), culminating in the rise of a new area of study named complex networks or network science (Bragin, 2010). Such networks can be found on many areas (Costa et al., 2007), from computer systems to protein interactions and, of course, in social networks. The creation of OSNs and the availability of data, thus, are leveraging this emergent study of complex attributes of OSNs.

4.1 Topology characterization

Analysing the topology of a social network can reveal several interesting features about its components and how people organize themselves for different purposes. Extracting network connections from OSNs is much easier than in offline networks, as all required data is already stored digitally, not asking for explicit knowledge extraction strategies.

Several SNSs had their networks explored and their topologies characterized, such as (to name a few):

- General OSNs services – Facebook (Kumar, 2012), Orkut[28] (Ahn et al., 2007; Mislove et al., 2007), Myspace (Ahn et al., 2007), Cyworld[29] (Ahn et al., 2007; Chun et al., 2008);

- Media sharing services – YouTube, Flickr (Mislove et al., 2007);

26. This is known as the small-world effect, in which the average distance between nodes increases slowly (proportional to $\log N$) in relation to the number N of nodes in the network.
27. In a power-law distribution, the probability of a node to have degree (number of connections) k is given by $p(k) \propto k^{-\gamma}$, where γ is a positive constant.
28. `http://www.orkut.com` (defunct since September 2014)
29. `http://global.cyworld.com` (defunct since February 2014)

- Blogging services – Twitter (Huberman et al., 2008; Kwak et al., 2010), LiveJournal[30] (Mislove et al., 2007);
- Message exchange services – MSN messenger (Leskovec and Horvitz, 2008).

In addition to these services, some studies also attempted to characterize the topology of social networks formed implicitly in sites like university web pages (Adamic and Adar, 2003) and email groups (Adamic and Adar, 2005; Tyler et al., 2005).

What the network structure reveals

One important property revealed by topology characterization is how similar OSNs are to other real networks previously studied. Agreeing to what is observed in offline social networks, Mislove et al. (2007) verified the presence of power-law degree distribution and small-world property in several OSNs. Kwak et al. (2010) discovered, however, that Twitter's structure does not follow a power-law degree distribution, having an unusual high number of popular users with many followers[31], therefore resembling more a news network than a social network.

Using data from MSN Messenger, Leskovec and Horvitz (2008) analysed the mean distance between users, identifying small-world property in this network and also showing how people with similar interests (same age, language, location and opposite sex) tend to connect and keep frequent communication. Kumar (2012) discovered that 99.91% of Facebook users belong to the same large *connected component*[32] and that friends communities[33] can be stunningly dense, compared to the general sparse structure of the whole network. Also, they showed that common age and nationality are relevant to determine social connections.

The network characterization in services where there is no explicit network allows the inference of interesting discoveries. By characterizing the network formed by internal links connecting web-pages from a university domain, Adamic and Adar (2003) showed possibilities of discovering communities and real-world connections among students. From networks built from email services, Tyler et al. (2005) were able to perceive hidden patterns of collaboration and leadership among users, identifying communities (formal and informal) and leadership roles within the communities.

Many networks in one network

An interesting fact is that an OSN may embed more than one network structure. Many SNSs explicitly register users' relationships, resulting in a *friendship network*. However, from users' interactions, an implicit *interaction network* can also be formed, revealing which social connections are actually active and in use (generally a subgraph of the friendship network).

30. `http://www.livejournal.com`
31. On the Twitter network, connections between users are directional, where one side of a connection is a "follower" and the other a "followee". Followers receive all the contents posted by the followees, while the reverse is not necessarily true.
32. In a network's connected component, there is a path between each pair of nodes belonging to it. In practice, a huge connected component, like the one found on Facebook, means that almost all users in the network can be reached by any other user in Facebook using only existing social connections.
33. Communities of users can be defined either explicitly, in SNSs where users declare membership to specific groups, or implicitly, as a topological property of the network (which is the case of the article cited here). A topological community is defined by a group of users strongly connected among them, but weakly connected with other groups (Girvan and Newman, 2002).

Other possible implicit networks are *diffusion networks*, characterized by the course of a content in the network, and *interest networks*, defined by groups of people with similar interests.

By comparing the friendship network to the interaction network on Twitter, Huberman et al. (2008) showed how smaller is the second one, but more adequate to describe and analyse social events. Chun et al. (2008) showed how Cyworld's interaction network can be more precise to represent real networks, having its nodes' degree distribution closer to known social networks than the friendship network. Wilson et al. (2009) discussed that the interaction network can present a different perspective and metrics for an OSN (like larger network diameter and less connected "supernodes"), being suitable for applications like social spam detection and online fraud detection.

Smith et al. (2014) analysed conversations on Twitter about different topics and identified, from how participants of a topic are connected, the formation of six distinct network structures according to the subject being discussed. These network structures describe different "spaces" of information exchange: from the engaged and intransigent crowds, to the fast content replicating and sharing broadcast networks.

4.2 Use and functionality characterization

Since the rise of SNSs, researchers have been interested in understanding the functionality of those services and how their users could take advantage of them.

Network formation
While SNSs were still becoming popular, Backstrom et al. (2006) described how the OSN structure can impact in new friendships and community formation. They showed that more densely connected communities are more likely to receive new members and that events, as the change of the topics of interest in a group, tend to cause transformations in the network topology. Wilkinson (2008) made similar discussion, but focusing on networks of peer production services (Wikipedia[34], Digg, Bugzilla[35] and Essembly[36]), showing how more ancient individuals have a tendency of receiving new connections, concentrating contributions and remaining longer in the network.

Java et al. (2007) described, in an introductory perspective, what is Twitter and the main uses of the service: talking about everyday subjects and finding information. Then, they showed how coherent communities arise from the aggregation of users with similar interest. Takhteyev et al. (2012) analysed how users' geographical distribution affects their links, uncovering a correlation between the existence of a connection among two users and the frequency of airline flights between the cities they live.

User profiles
Network users can be categorized in different classes by their attributes and patterns of behaviour. Krishnamurthy et al. (2008) analysed profiles of almost 100,000 Twitter users and identified three different classes of users: *broadcasters*, with much more followers than followees (e.g.: celebrities); *acquaintances*, with reciprocity in their relationships (e.g.: ca-

34. `http://www.wikipedia.org`
35. `http://www.bugzilla.org`
36. `http://www.essembly.com` (defunct since May 2010)

sual users); *miscreants*, that follow a much larger number of users than they are followed (e.g.: spammers or stalkers).

Wu et al. (2011) identified "elite" Twitter users (i.e., celebrities, famous bloggers, media and corporation accounts) and evaluated the impact of the content published by them, realising that half of the URLs that circulate over the network are generated by 20,000 of those "elite" users. Association patterns among those special users are also analysed, revealing that "elite" users of a same field (e.g.: celebrities or blogs) tend to interact among them.

Benevenuto et al. (2009) were able to analyse and measure the online activity of users of four SNSs: Orkut, Myspace, hi5[37] and LinkedIn[38]. They discovered that users spend on average 92% of their time on those services just browsing other users' pages, without posting any content to the network.

Conversation

A notable feature of OSNs is the users' ability to maintain conversations, enabling the organization of collective mobilization and the creation of enriched content. Kumar et al. (2010a) elaborated a detailed study of how conversations are created in diverse OSN contexts, finding patterns and particularities that enabled the creation of a simple mathematical model capable of describing the dynamics of the conversations.

Honeycutt and Herring (2009) analysed how conversation dynamics can occur on Twitter, with users adapting its simple mechanism of message exchange to track and maintain active communication with each other. In the same line, Boyd et al. (2010) explored how *retweets*[39] can be used to create conversations and involve new users in existing conversations.

Discussing the impact of communication in OSNs, Bernstein et al. (2013) discovered, by analysing large amount of log data, the extent of diffusion of content published on Facebook (i.e., how many people read a message posted by a user). They showed that users usually underestimate the extent of their posts, expecting an audience of less than one third of the actual reached audience.

Network deterioration

Not only the growth, but also the decline in the use of SNSs was studied. Kwak et al. (2011) examined details of the *unfriending* (i.e., unfollowing) behaviour on Twitter, showing how frequent it is, using both quantitative and qualitative data, which were obtained through user interviews. Garcia et al. (2013) examined SNSs that suffered intense decline in user activity (Friendster[40] and LiveJournal), attempting to understand the impact of users desertion. The impact of "cascades of users leaving" on the network resilience was deeply studied, and a metric was proposed to determine when it is or it is not advantageous to users to join a network.

37. `http://www.hi5.com`
38. `https://www.linkedin.com`
39. A *retweet* is a common practice on Twitter, where a user reposts a message (*tweet*) previously posted by another user, commonly as sign of support or reinforcement.
40. `http://www.friendster.com`

4.3 Representation models

One of the challenges of OSN studies is to create models able to describe with success the structure, events and transformations the network goes through. Different models have been proposed addressing this issue. We discuss some of them below.

Structure models
When analysing the structure of photo sharing OSNs (Flickr and Yahoo!360[41]), Kumar et al. (2010b) detected patterns in the network representing different regions: *singletons* (users without connections), *isolated communities* (generally around a popular user) and a *giant component* (users connected to many users). Then, a simple generative model was proposed, able to reproduce the network evolution and recreate the structural patterns observed empirically.

Xiang et al. (2010) worked on building a model able to represent the *intensity* of social relationships. Instead of having a binary value, each edge between two users in a social graph is calculated as a function of the frequency of interaction among them.

Spatio-temporal models
Although graphs are suitable representations to analyse spatial properties of an OSN, temporal aspects must also be considered in order to represent transformation processes taking place in a network. Although observing temporal aspects of OSN can be a challenge (specially due to the huge number of users involved in processes and data retrieval restrictions), they can be a valuable source of information.

The temporal evolution of a network was studied by Leskovec et al. (2005), who were able to make interesting empirical observations about the growth of several real networks. They noticed that, contrarily to the expectations, the addition of new nodes makes the network become denser in terms of edges per nodes and the average distance between nodes often decreases over time. From those observations, a graph generator model was proposed, able to produce more realistic networks.

Tang et al. (2010) proposed temporal models to describe network transformations, enabling the creation of new metrics, like temporal distance, i.e., the average time taken for an information published by a user to reach other users. Those metrics are complementary to other spatial metrics (such as geodesic distance) and seem to enable new perspectives of analysis of information diffusion processes or network formation.

5. Social data analysis

The focus of social data analysis is essentially the *content that is being produced by users*. The data produced in social networks are rich, diverse and abundant, which makes them a relevant source for data science. As will be seen in this section, most of the computational researches that employ social data use it in machine learning problems such as natural language processing (NLP), classification and prediction. In addition to the challenge of building robust algorithms for such purposes, researchers have also the challenge of building scalable computational solutions that can deal with the large amount of data available in those services.

41. `http://360.yahoo.com` (defunct since July 2009)

5.1 Sentiment analysis

The textual information produced everyday in SNSs, like Twitter, is a huge corpora (Pak and Paroubek, 2010), in which natural language processing techniques, such as sentiment analysis, can be used. Applied to OSNs, sentiment analysis has the potential to describe how emotions spread among populations and their effects.

Taking advantage of corpora particularities
New sentiment classification strategies can be explored, if particularities of the services are taken into account, like the Twitter's (short) size of messages, slangs, *hashtags*[42] and network characteristics. Nichols and Fisher (2007) is one of the first attempts in the literature of sentiment classification on Twitter. Text processing techniques were proposed to extract and reduce features and an algorithm was built reaching over 80% of accuracy in classification. Hu et al. (2013a) also noted that an interesting feature of social data is the presence of *emoticons*, that can be used as labels for machine learning algorithms, helping the process of classification.

Another interesting element of OSN corpora is the presence of language expressions not always present in formal texts. Using the fact that sentences are commonly followed by descriptive *hashtags* (like "#irony" or "#not") that can be used as labels for supervised learning, Culotta (2010a) and Reyes et al. (2012) worked on learning and detecting sarcasm and irony in text, with positive results.

Applications
Sentiment analysis can have many applications. For example, Jansen et al. (2009) and Ghiassi et al. (2013) analysed how OSN users express sentiments towards different brands, obtaining a measure of approval or disapproval. With the increasing influence of SNSs, this kind of work can be valuable for companies to understand and deal with customer demands. Dodds et al. (2011) and Lansdall-Welfare et al. (2012) developed indicators of happiness among populations, based on the analysis of OSNs texts. With that, they were able to analyse the impact of historical events—such as economic recession (Lansdall-Welfare et al., 2012)—in public opinion, showing an innovative quantification of population welfare.

Deeper analyses take into account not only text classification, but also a study of how sentiment spread in the network. Hu et al. (2013b) took advantage of *emotional contagion* theories (Howard and Gengler, 2001) to help the classification of texts produced by specific users, having better results than traditional algorithms. In a recent (and controversial) experiment, Kramer et al. (2014) filtered content displayed on Facebook to emphasize positive or negative posts, showing how emotions can be contagious. Although the experiment did not produce expressive change in users behaviour, the observed effects were statistically significant.

5.2 Prediction

A valid question to address when dealing with OSNs is how representative are the dynamics present in the virtual environment in relation to the non-virtual world. Supposing that what

42. Hashtag is a text prefixed with the hash (#) symbol. It is commonly used in SNSs to label or tag messages.

happens inside SNSs can provide information about other external events, researchers have been trying to build predictors in many fields:

- Elections: predict the outcome of elections from OSNs manifestations (Tumasjan et al., 2010);
- Box-office revenue: forecast the popularity (and revenue) of a blockbuster before or just after it comes out (Asur and Huberman, 2010);
- Book sales prediction (Gruhl et al., 2005);
- Disease spread (Culotta, 2010b; Lampos and Cristianini, 2012);
- Stock market prediction from sentiment analysis (Bollen et al., 2011b).

However, despite the initial positive results and good perspective presented in the works above, skepticism about the effectiveness of the proposed methods and their representativeness must be noted, as seen in Gayo-Avello (2013), Wong et al. (2012) and Zhang et al. (2011), which analysed election forecasts, box-office revenue and stock market predictions, respectively. Those studies showed that the validity of the initial findings can be questioned and that many results can not be generalized as expected.

5.3 Trending topics detection

Another important focus of research that uses content published in OSNs is the analysis of message exchange dynamics, aiming to detect trends. Although some SNSs, like Twitter, have their own algorithms for trending topics detection, alternative proposals of content detection and organization have been made. According to Guille et al. (2013), there are two main approaches to detect a trending topic in an SNS: message analysis or network analysis.

Message analysis
Focusing on the messages content, Shamma et al. (2011) proposed a simple metric to identify trending topics, analysing the frequency of words during specific time frames, compared to its general frequency (similar to the usual tf-idf (Dillon, 1983) model in NLP). A trending topic happens when there is an abnormal term frequency occurrence. In a creative approach, Weng et al. (2011) considered the frequency in time of words as waves. When some phrase's terms have resonance among them, emergent topics are identified.

Lu and Yang (2012) went beyond and developed a method to predict which topics will be popular in the future. Using strategy originally intended to predict stock markets, this method is able to calculate the *trend momentum*: the difference of frequency of a term between a short and a long time period. In the tests performed, the method was effective, with trends being successfully predicted by the increase of the momentum.

Network analysis
On the transition from message to a network approach, Cataldi et al. (2010) used not only the term frequency, but also the authority (calculated using PageRank) of users posting the observed content. This way, they were able not only to identify trending topics, but

also related topics. Takahashi et al. (2014) used exclusively network information to create a probabilistic model of interactions. When anomalies are detected in the interaction pattern, a trending topic can be detected, without even text analysis. In their tests, this technique performed at least as good as other text-based techniques, being superior when topic keywords are hard to determine.

Tracking memes evolution
Apart from trending topics detection, Leskovec et al. (2009) studied not only topics created, but also their evolution in new subtopics or derivatives over time, observing the spreading of news for days. The researchers were able to track a common path in the news cycle, with content being first published in traditional media and, few hours later, the same content appearing in blogs and other online services, resulting in "heartbeat-like" patterns of attention peaks.

In the same line of meme detection, Kuhn et al. (2014) analysed the evolution of scientific ideas, identifying repeated terms and the connections in the network of citations between scientific papers. Thus, they developed a "meme score" from terms' frequency and propagation degree, enabling the identification of relevant concepts within a scientific field.

5.4 Real event detection

In many cases, topics discussed in OSNs are about events that take place in the "real" (or external) world, like political or public events. Also, as contents are often posted from mobile devices, it is common for OSN users to be physically present during those events. As many SNSs aggregate *geolocation* tags to such posts, OSN data can allow access to the place from where they were sent.

The geolocation of an OSN user can have influence on relationships and content exchanged, as shown by Takhteyev et al. (2012). Cheng et al. (2010) indicated that it is possible to predict the location of a user exclusively from the content of his/her textual messages, even when this information is not explicitly disclosed.

Detecting real events
Becker et al. (2011) worked on a method to distinguish Twitter messages that refer to real events from those that do not (jokes, spam, memes, etc.) by clustering messages of the same topic and, then, classifying the clusters based on their properties. Psallidas et al. (2013) discussed the challenge of separating, in an OSN, content related to predictable events (e.g.: awards, games, concerts) from those related to unpredictable ones (e.g.: emergencies, disasters, breaking news). Features useful to describe each type of diffusion were evaluated to be used as input to classification algorithms, being effective in large-scale experiments.

Sasahara et al. (2013) analysed how some topics related to past events spread across the social network, finding some patterns that help in the identification of real event diffusion. According to the authors, diffusion networks of real events have an abrupt and unusual structure (compared to diffusion of other kinds of events), making it possible to create automatic tools to detect them.

Using real events information
Hu et al. (2012) studied how a social network is capable of disclosing breaking news even before traditional media. They used as case study the fact that the news of Osama Bin

Laden's death were disclosed on OSNs before traditional media and showed how OSN users take roles of leadership to efficiently transmit information and influence other users on those events.

Using this ability of quick awareness, Sakaki et al. (2010) and Neubig et al. (2011) proposed automatic methods for detecting earthquakes in Japan, considering network users as "social sensors". Their results were robust and promising, involving the identification of the earthquake's center and trajectory, inference about the safety of people possibly affected and the generation of automatic earthquake alerts faster than official announcement by authorities.

5.5 Social recommendation systems

Another application of OSN data is the possibility of creating social recommendation systems for products or even content produced by users in the network. In a space with many users and data, the use of social relationships can improve traditional recommendation systems both in relevance and scalability, as users connected by social relationship usually share many interests, both by homophily[43] and by contagion, reducing the amount of data necessary to make accurate recommendations.

Trust networks

One practical use of social information in recommendation systems is the synthesis of *trust networks*, which are groups of related users that are considered to have a valuable opinion on some matters. Generally, a user's truthfulness is related to its proximity to a reference user.

Walter et al. (2007) described how an OSN can be used to collect information in general and how the relationships can help to filter relevant information for each user, as trust networks are established. By using exclusively content on users' neighbourhoods, they were able to build effective recommendation systems as good as other systems that use information from the whole database. Arazy et al. (2009) created social recommendation systems in order to evaluate products reputation, building trust networks to ponder the relevance of users opinions.

Improving traditional recommendation systems

Other uses of OSN data for recommendation systems include the work of Ma et al. (2011), who uses relationship data to initialize recommendation systems that have few initial reviews. Also, Yang et al. (2013) created probabilistic models to model users preferences and make recommendations based on friendship connections. In a more conservative proposal, Liu and Lee (2010) suggested ways to improve existing recommendation systems by including social information, like users' relationships, and showed how the accuracy of algorithms may be positively affected.

Content selection

A common task for recommendation systems on SNSs is to select relevant content to be displayed to users. Chen et al. (2010b) worked on a series of algorithms to recommend content to users, in order to improve Twitter's usability. They were able to reach a level

43. The tendency of an OSN user to connect to similar people.

in which 72% of the content showed was considered interesting, according to real Twitter users feedback.

Backstrom et al. (2013) worked with Facebook data, analysing the attention a topic might receive, by predicting the topic's length and its re-entry rate (i.e., the number of times a user participates in the same topic). This gives a measure of how interesting a topic is and can be used to select and recommend content to users.

6. Social interaction analysis

By watching *users diffusing content*, there is the expectation of knowing more about complex human behaviour. The access to data produced by OSNs and the knowledge of how to process and analyse them are enabling computer scientists to join discussions previously exclusive to sociologists or psychologists. This new intersection of fields is known as *computational social science* (Lazer et al., 2009; Cioffi-Revilla, 2010; Conte et al., 2012).

There are still questioning related to whether the behaviour observed in an OSN can be extrapolated to its users offline lives and whether OSN users are representative enough for drawing conclusions, from their behaviour, for whole societies (Boyd, 2010). Even so, there is a plenty of phenomena that take place on OSNs that are worth to be studied, as we will outline in this section.

6.1 Cascading

One of the most widely studied behavioral phenomenon that takes place in OSNs is information cascade. Also known as viral effect, a cascade is characterized by a contagious process in which users, after having contact with a content or a behaviour, reproduce it and influence new users to do the same. This decentralized process often causes chain reactions with great proportions, involving many users and being one of the main strategies for information diffusion in social networks.

The unpredictability and the magnitude of this phenomenon attract many researchers, trying to interpret and understand the factors behind it. The cascade effect has been studied and characterized in many different SNSs, as:

- Facebook (Sun et al., 2009; Dow and Friggeri, 2013);
- Google+ (Guerini et al., 2013);
- Second Life[44] (Bakshy et al., 2009);
- Flickr (Cha et al., 2009);
- Twitter and Digg (Lerman and Ghosh, 2010).

Goel et al. (2012) alone studied information diffusion in seven different OSN domains, verifying similarity in cascading properties, regardless the service observed.

44. `http://secondlife.com`

Properties observed
From the empirical analysis of information cascades on OSNs, some common properties can be observed, as already shown by Goel et al. (2012). A good characterization of many of those properties can be found in Borge-Holthoefer et al. (2013), that gathered results from works that modeled and analysed cascades.

Among the properties observed, some are highlighted:

- Most cascades have small depth[45], exhibiting a star-shaped connection graph (a central node connected to many others around it). This was shown by many researchers, as Leskovec et al. (2007), González-Bailón et al. (2011), Lerman and Ghosh (2010) and Goel et al. (2012).

- The majority of information diffusion processes that take place in the network are shallow and do not reach many users. Thus, widely scattered cascades turn to be rare and exceptional events.

- In general, cascades (even large ones) occur in a short period of time. Most reactions to a content posted on an OSN usually happen quickly after it is posted (Centola, 2010; Leskovec et al., 2007) and do not last for a long time (Borge-Holthoefer et al., 2013).

- Any user on the network has potential to start widely scattered cascades. It is shown that different sources of information can conquer space on the network (Bessi et al., 2014), and attempts to measure users' potential to start a cascade are not conclusive (Bakshy et al., 2011; Borge-Holthoefer et al., 2012) (see section 6.5 on influence for more details).

Information origins
Myers et al. (2012) studied sources of information in OSNs. They found that almost one third of the information that travels on Twitter network comes directly from external sources, while the rest comes from other users, through cascades. Tracking a cascading process can be a challenge when the content being propagated may undergo changes. Leskovec et al. (2009) proposed ways to track memes and their derivatives, in a process that can take several days, showing the long transformation process from publication to popularization.

How topology influences cascades
The analysis of the network underlying a diffusion is a helpful way to understand a cascading process. Goel et al. (2013), using a dataset of billions of diffusion events on Twitter, analysed the diffusion networks and proposed a "structural virality" metric, able to measure the network's tendency to successfully propagate an information.

One of the most important conclusions of the network analysis, shown by Sun et al. (2009), Ardon et al. (2013) and Weng et al. (2013), is the fact that topics that can reach initially more than one community of users tend to cause larger cascades.

45. The depth of a diffusion network (or tree) is the maximum distance between the diffusion source (the root) and the users involved in the diffusion. A distance between two users is defined as the size of the shortest path on the network that connect them.

Cascades from historical events
Specific events where SNSs had significant influence, such as political movements and protests, received special attention in social network analysis. In 2009, following the Iran presidential elections, many protests took place and their effects could be noticed in SNSs by increased diffused information. Zhou et al. (2010) conducted a qualitative research of these cascades, concluding that in general they are shallow (99% of the diffusion trees have depth smaller than three). González-Bailón et al. (2011), based on the diffusion network, analysed the roles of users and related them to their positions in the network. According to the study, influential users in the process of spreading information tend to be more central in the network.

Similar experiments were made with protests that happened in Spain on May 15th 2011. Borge-Holthoefer et al. (2011) analysed the diffusion network related to such events and differentiated users that acted as sources of information and users that only consumed it. In a later work, Gonzalez-Bailon et al. (2013) identified four types of users—namely influentials, hidden influentials, broadcasters and common users—that can help the understanding of how users behave in cascading processes.

6.2 Predicting cascades

An important motivation for characterizing cascades is to be able to predict how users in a network will behave with regards to a specific content and how this content will spread. This capacity to tell beforehand how many users will see or share an online content can be a source of revenue for advertisers and, also, a useful tool to governments willing to effectively disseminate public interest information.

However, the task of predicting popularity of online content has shown to be extremely difficult to accomplish (Salganik et al., 2006; Watts, 2012). Two main problems are determinant (Chen et al., 2010a): (a) the definition of what are the features (if any) that determine the size of a cascading process; and (b) the fact that widely spread cascades are rare events (Goel et al., 2012), making it hard to develop and train algorithms with so few positive samples.

Nevertheless, those difficulties were not enough to prevent research in this area, as seen in the many scientific works already published. Also, according to experiments presented by Petrovic et al. (2011), the identification of content likely to be shared is a task manageable by humans, what can bring hope to new inquires. As we will show below, many are the works published in this topic and so are the strategies used to tackle the problems.

Feature selection
The most important aspects to be considered when building machine learning algorithms (such as predictors or classifiers) to analyse cascades is the proper characterization of information diffusion processes and the choice of relevant properties to describe these processes preserving existing distinctions among them (Suh et al., 2010). From the literature, we can see that four main classes of features are generally chosen: (a) message features, (b) user features, (c) network features, and (d) temporal features.

Message features
Does a textual message posted in an OSN have an intrinsic potential to be shared? Assuming

that some content has more potential than others to create cascades, researchers have investigated ways of predicting the future popularity of a message based on text analysis. This kind of investigation might be specially interesting in cases in which there is the need (or the will) of maximizing the audience reached by a content posted by a specific user. Thus, by adjusting the text that will be posted, it would be possible to increase the range of an author's message.

This is the aim of Naveed et al. (2011) work, that found correlations between message content and *retweet* count on Twitter. Several features were analysed, such as presence of URLs, *hashtags*, mention to other users, punctuation and sentiment analysis. Their conclusion is that messages referring to public content and with negative emotions are more likely to be shared. Suh et al. (2010) did an extensive search for features, both in message and user characteristics, in a large dataset (74 million posts from Twitter) highlighting the presence of URLs and *hashtags* as the most relevant factors in the message content for predicting cascades.

More creative message descriptors were studied by Hong et al. (2011), who used topic detection algorithms to identify a message's topic, to be further used as a feature. Tsur and Rappoport (2012) explored different interesting features that can be extracted from a *hashtag*, like its location inside a post or its size in characters or words.

User features

It is evident that a popular and influential user has more chance of generating a cascading process than an anonymous user. Therefore, analysing aspects related to the user that shares a message, and possibly about the users that continue this process, can be crucial to build a reliable cascade predictor.

In addition to message features (as discussed above), Suh et al. (2010) also analysed a set of possible features related to authors, including the number of connections, number of past messages posted, number of days since the user's account was created and number of messages previously marked as favorite by other users. Their conclusion was that only the number of connections and the age of the account have any sort of correlation to *retweet* rates. Hong et al. (2011) also suggested other features, namely: author's authority according to PageRank (Page et al., 1998), degree distribution, local clustering coefficient[46] and reciprocal links.

Metrics taking into account properties of the users involved in a diffusion (beyond the author) can be also valuable. Hoang and Lim (2012) introduced a model to predict information virality on Twitter, by creating three features: item virality (the rate of users that share a content, after receiving it), user virality (the number of connections of users involved in a diffusion) and user susceptibility (the proportion of content shared in the past by a user). Hogg and Lerman (2009b), by observing cascades on Digg, were able to create models that describe the initial behaviour of users sharing content, thus allowing the forecast of a cascade's size. Lee et al. (2014) explored features related to previous behaviours of users, such as average time spent online, time of the day in which the user is more likely to join discussions, and number of messages sent over time.

46. Clustering coefficient is a measurement of network cohesiveness. The local clustering coefficient for a specific node is given by the number of direct connections between two of its neighbors, divided by the number of possible connections between these neighbors.

Network features
The analysis of the network structure where a diffusion takes place is also important to determine the potential range of a cascade.

Weng et al. (2013) explored the importance of a network characterization, using the knowledge that diffusions starting in multiple communities are more likely to be larger (Sun et al., 2009; Ardon et al., 2013). The authors then proposed as a metric the number of communities involved in the early diffusion and the amount of message exchanges between different communities (inter-community communication).

Kupavskii et al. (2012) examined a set of features to describe a cascade, showing relevant improvements in the prediction task when using network features such as the flow of the cascade—a measure related to the number of users sharing a content and how fast they share it—and the authority in the network formed by users sharing the same message, calculated using PageRank (Page et al., 1998). Ma et al. (2013) used both message and network features to predict the popularity of Twitter *hashtags*. Among the network features adopted are metrics like the ratio between the number of connected components in a network and the number of users that initiated the cascade, the density of the diffusion network[47] and the diffusion network's clustering coefficient. Their conclusion is that network features are more effective than message features for predicting the use of *hashtags*.

Temporal features
Every cascade process can be represented as a time series, listing the amount of information diffused over time. This time series can be seen as a cascade signature, representing its range, speed and power.

Szabo and Huberman (2010) analysed the initial diffusion of YouTube and Digg contents and, based on the initial time series, forecast the long term popularity of specific contents. It wad realized that only two hours of data about the access to Digg stories was enough to predict thirty days of popularity, while, on YouTube, ten days of records were needed to evaluate the next twenty days.

Chen et al. (2010a) improved this strategy, by dividing the original prediction problem into subtasks where, based on past features, a classifier must estimate if a content published on Facebook will double its audience or not. Thus, robust and high performance classifiers can be built.

What exactly is predicted
After presenting the features used to describe cascading phenomena, it is worth examining the different approaches to predict cascades.

Most of the work in this topic tries to measure the number of users or messages that will join a cascade. Examples are Kupavskii et al. (2012), who worked predicting the number of messages (*retweets*) a cascade will have, Ma et al. (2013), that predicted the popularity of a new topic (*hashtag*), and Suh et al. (2010), that forecast the rate of users participating in a cascade.

However, some works were simply interested in building binary classifiers to determine if a content will be shared by any user or not. This is the case of Naveed et al. (2011) and Petrovic et al. (2011). Hong et al. (2011) went a little further and created four

47. The density of a network is the ratio between the number of actual connections and the number of possible connections.

categories of cascading—not shared, less than 100 shares, less than 10000 shares and above 10000 shares—that can be classified more easily.

Another strategy was used by MORGAN (2009), who built a system able to predict which users are leaned to enter a cascade. Lee et al. (2014) worked in the same line, being able to sort the N users most inclined to share a message.

6.3 Rumours diffusion

Another particular area of study involving cascading that received special attention from the research community is the detection of false information (rumour) propagation.

Characterizing rumours
Aiming to characterize this phenomena, Friggeri et al. (2014), with the assistance of a website that documents memes and urban legends (`http://snopes.com`), mapped the appearance of rumours on Facebook network, showing that rumour cascades tend to be more popular than generally expected and discussing users' reactions after acknowledging the falsehood of previously posted messages. Also on Facebook, Bessi et al. (2014) observed the acceptance by network users of different sources of information. By analysing how content from (a) mainstream media, (b) alternative media, and (c) political activism is diffused, they concluded that, regardless of source, every information has the same visibility. This may favour people that share false content, as they potentially have the same power of influence on the network as reliable sources.

Detecting rumours
Mendoza et al. (2010), when analysing the diffusion networks of news related to a natural disaster in Chile, realized that the patterns of rumour spreading are different from those related to real information spreading. Therefore, in a subsequent work, Castillo et al. (2011) sought automated methods to detect rumors, by analysing features from texts posted and the users involved in the propagation of the information.

Qazvinian et al. (2011) further proved the effectiveness of using features related to network and message content to detect rumours. Despite their positive result, it is noticeable the small number of rumours analysed (only five), given the quantity of data (10000 posts from Twitter), raising doubts about the method's generalization power. Gupta et al. (2012) also worked developing metrics, but this time trying to measure credibility of users, messages and events, resulting in a score for the credibility of the general topic diffused.

Rumour containment
In a different perspective, Tripathy et al. (2010) explored ways to contain a rumour cascade, after its identification. Using techniques inspired by disease immunization, they discussed the importance of a quick identification of rumours and the use of anti-rumours agents able to detect such events and spread messages against the rumours. Lastly, Shah and Zaman (2011) aimed to detect the source of a rumour cascade, developing a new topological measure entitled "rumour centrality", able to outperform traditional metrics in special cases.

6.4 Information diffusion models

One way to understand and study the dynamics of OSNs is to build models that represent users interactions. Having a reliable representation enables the conception of simulations that can give support to understand the events that take place in the network.

Models paradigms

In Borge-Holthoefer et al. (2013), the models used to describe cascades in complex networks are revised. According to them, the models can be divided in two main groups: (a) threshold models and (b) epidemic and rumour models. In both methods, the decision of a user to adopt a certain behaviour depends on the neighbours that have already adopted it. In threshold models a user will act only if the proportion of his/her neighbours that are active is superior than a given threshold; in epidemic and rumour models, on the other hand, active users have a probability of infecting each of their neighbours.

An example of the threshold model is provided by Shakarian et al. (2013), using the model to create a heuristic to identify users able to start a cascade. The method is able to quickly identify a relatively small set of users able to start cascades that cover the whole network, even for huge networks with millions of nodes and edges.

Using the epidemic model, we have the work of Gruhl et al. (2004) who created a model for information diffusion in blogs, using real data to validate it. It was shown that the model faithfully reproduces real behaviour, where influential and popular blogs in reality also have relevance in the model's diffusion. Golub and Jackson (2010) also showed that the epidemic model is an appropriate form of representing cascades, when modeling (the rare[48]) high depth cascades.

It is important to notice that the epidemic model, based on disease propagation, has its limitations when describing information contagion, given their different nature. One important distinction is the concept of *complex contagion* (Centola and Macy, 2007) which states that, for a behaviour be acquired by an individual on social networks, he/she has to be exposed to multiple other individuals. This differs from disease infections, where a single contact with a virus is enough to infect a person (*simple contagion*). Romero et al. (2011a) explored this phenomenon on Twitter, showing that multiple exposure to subjects were determinant for contagion. Weng et al. (2013), however, made a counterpoint showing that although most content spread like complex contagion, some can be properly modeled as simple contagion.

In a different approach, Herd et al. (2014) built a model where, after collecting behaviour data from Twitter, each user receives a probability of posting and a probability for emotions to be expressed. With this, they created a multi-agent model to simulate the behaviour of social networks. By building a model based on messages exchanged during United States 2012 presidential campaign, the researchers were able to detect which users were more influential to spread messages. An unexpected conclusion was the fact that the removal of the ten biggest enthusiasts of Barack Obama's campaign would have a larger impact in the network than if Obama himself was removed.

48. As noted before, most cascades observed empirically present small depth. However, in Liben-Nowell and Kleinberg (2008), "large and narrow" diffusion trees were observed (probably due to the nature of the content being observed—email chains—and to the set examined—successfully diffused chain letters) and were taken as the base structure used on the work of Golub and Jackson (2010).

Model enhancements
Some enhancements can be proposed to turn the models more realistic to the OSN context. This is the case of Weng et al. (2012) and Gonçalves et al. (2011), which considered limitations on the amount of information each user can access and process. This is able to reproduce the fact that many of the information diffusion on OSNs simply lose strength and disappear, regardless the content.

Gómez et al. (2013) discussed ways of modeling and processing information diffusion through multiplex networks. A multiplex network is a network with multiple levels, each level representing a different type of relationship between the network nodes. Therefore, a multiplex network is an adequate model to represent online social networks, as OSN users can be connected in multiple ways (e.g.: different topics may generate different dynamics on the network, creating different diffusion networks connecting users). The proposed analysis revealed relevant aspects of the relationship among those multiple processes.

Inferred paths of propagation
Another area of interest is to determine which are the paths traveled by messages subject to diffusion. Gomez-Rodriguez et al. (2012) were able to infer the order in which users were "infected" by a content, by observing the final infected network. By analysing the timestamps when network nodes shared a content, they calculated the most likely structure that connects the nodes. The algorithm is applied to a large database of blogs' diffusions, achieving high quality results.

Yang and Leskovec (2010) created a method to model and forecast information diffusion, independently of the network structure. For each user of the network, an influence index is estimated, as a measure of the number of users infected by him/her, over time. Thus, for an initial group of infected users, it is possible to predict how many new users will be infected in the future, even without information regarding their connections. Also, the individual influences can be grouped and be used to model the influence dynamics of different classes of users.

6.5 Influence

As already antecipated, another important factor that determines information diffusion in an OSN is the users' capability of influence. An influential user can be determinant to start (or trigger) cascade events, or even change people's opinion and behaviour.

Locating influential users
Locating an influential individual in a network is not a trivial task. Cha et al. (2010) discussed three metrics aiming to quantify users' influence in OSNs: number of connections (nodes degree), number of mentions, and number of messages reshared (*retweets*) by other users. A discussion of the most appropriate ways to measure influence is done, revealing that simple metrics like number of connections can be misleading to represent the future influence of a user. Weng et al. (2010) were more optimistic, showing that an adaptation of the PageRank algorithm (Page et al., 1998) can be used to successfully measure influence on networks.

However, Bakshy et al. (2011), when analysing a huge dataset, showed that the theoretical results and metrics are not always confirmed in reality. They discussed that, even though

it is possible to identify influential users able to repeatedly start widely scattered cascades, determining a priori which users will influence a cascade process is a hard task. Borge-Holthoefer et al. (2012) also analysed real data in order to identify influential users from the network topology. Although some influential users are correctly identified in some cases, there are situations where "badly located" users are also able to be influential, exceeding expectations.

Influence effects
Researchers have also been interested in evaluating the effects of social influence. Bakshy et al. (2009), by examining the adoption rate of user-to-user content transfer in Second Life[49] among friends and strangers, showed that content sharing among known users usually happens sooner than among strangers, although transactions with strangers can influence and reach a wider audience.

Stieglitz and Dang-Xuan (2012) analysed tweets with political opinions and concluded that texts with increased emotional words have stronger influence in the network, being more likely to be shared. Salathé et al. (2013) discussed how the network connections influence opinions and individual sentiment, by observing reactions to a new vaccine campaign in United States. It was shown that negative users are more accepted by the network and that users connected with opinionated neighbours tend to be discouraged from expressing opinions.

6.6 Network's influence on behaviour

Even though individual users have autonomy, it would be frivolous to assume that social connections do not have influence on the formation and evolution of their behaviours and opinions. The OSN analysis enables the empirical observation of the consequences of social connections on individual behaviour, and the development of new models and theories capable of explaining those hypothetical associations.

Homophily
A relationship between the topological structure of an OSN and the behaviour of its users can be often noticed. In most cases it is not possible to determine what is cause and what is consequence (i.e., if the topology is a result of users behaviour, or if the behaviour is a consequence of the topology), but the study of one can help in the understanding of the other.

Researchers identified, in general social networks, a tendency that users with common interests are usually connected to each other (McPherson et al., 2001). Such phenomenon is called *homophily* and is also verified on OSNs. For example, Bollen et al. (2011a) verified, by investigating the relationship between emotions and social connections, that users considered happy tend to be linked to each other.

Romero et al. (2011b) investigated the relationship between the (explicit) network of friendship and the (implicit) network of topical affiliations (i.e., the communities formed by users interested in a common topic). They showed that both networks have considerable intersection (users tend to connect to other users with common interests), such that it is

49. On Second Life's virtual world, users are able to share *assets* with other users. An asset can be an ability (e.g.: a dance movement), an item or other customizations.

possible to predict friendship from *hashtag* diffusions and also the future popularity of a *hashtag* from the friends network.

Users' information processing capability
Gonçalves et al. (2011) verified whether users are able to surpass, in OSNs, the *Dunbar's number*[50], given that users usually have hundreds, or even thousands, of connections in such services. After analysing message exchanges, they showed that, despite the abundance of social connections in OSNs, users are unable to interact regularly with more peers than what is predicted by Dunbar's threshold. Grabowicz et al. (2012) studied how the topology affects the type of content transmitted on the network, discussing how users not very close related (intermediary ties) can filter relevant information from several groups, while close relationships (strong ties) can be distracted with a great amount of irrelevant messages.

Divergence of opinions in networks
By examining the information diffusion dynamics on OSN, Romero et al. (2011a) studied how users would not immediately adopt an opinion or behaviour (such as a new political position) from the first contact with the idea, provided by few initial users. However, if the user is continuously exposed to such content, with many users reinforcing it, the chance of adoption increases. This result is validated on Twitter, where the authors examined how *hashtags* are diffused and the decisive role of multiple exposures.

Based on the relationships established on Twitter, Golbeck and Hansen (2014) estimated the political preferences of users and analysed how different political opinions coexist in a social network. Also, using the user database together with the predicted political preferences, they were able to analyse the audience of traditional media sources, classifying them as liberal or conservative. This media classification showed to be coherent with previous classification in the literature.

6.7 Self-organization

Some research groups studied how users in OSNs, given the absence of central command and their decentralized communication, are able to self-organize in specific situations.

Crisis events
Leysa Palen, Kate Starbird and colleagues (Vieweg et al., 2010; Starbird et al., 2010; Starbird and Palen, 2010, 2011, 2012) made a deep research on how OSNs can help managing information during crisis events, such as popular uprisings, political protests, natural disasters and humanitarian aid missions. The researchers identified that, among thousands of messages and publications during a crisis, there is the emergence of mechanisms able to deal efficiently with this overload of information. Some of the observed dynamics include the ability of content selection, relevance detection and attribution of roles to specific users. They showed that the largest information cascades during those events tend to happen with important content, being a way to emphasize content worth to be viewed by other users. Also, the network is able to identify reliable users (like on-site witnesses) and give relevance

50. The Dunbar's number is a limit, proposed by the anthropologist Robin Dunbar, for the maximum amount of stable social relationships one person is able to maintain. The actual number usually varies between 100 and 200 and was proposed based on observations of the relation between social group size and brain size in primates (Dunbar, 1992).

to their posts, by sharing them more often. Thereby, just by observing the content circulating on SNSs, it is possible to quickly identify the most important or urgent information and even coordinate actions in order to help and assist people.

Social curating

Another self-organizing ability of OSNs is content curating, which is the ability of collectively selecting and filtering content relevant to users. This process can happen both spontaneously in traditional SNSs or in dedicated services like *Pinterest*[51] or *Tumblr*[52], where users can collaboratively build collections of diverse subjects, selecting content from the Internet.

Liu (2010) explored the skills involved in the curating process, describing seven distinct abilities of a social network, namely: collecting, organizing, preserving, filtering, crafting a story, displaying and facilitating discussions. Those skills are compared to actual professional skills (archivist, librarian, preservationist, editor, storyteller, exhibitor, docent, respectively), emphasizing how impressive is the network ability to promote self-organization, being able to specialize and accomplish complex tasks.

Zhong et al. (2013), in a comprehensive study, described with details the process and the mechanisms of curating in Last.fm[53] and Pinterest services, discussing users motivations behind it. They also showed that the social curating process is able to give value to items differently from centralized strategies, being an important source of opinion and measurement of quality. However, the community choices can still be biased, specially when dealing with items already popular in the network, or previously promoted by the service.

7. Final remarks

In this work, we performed a comprehensive analysis of research published on online social network analysis, from a Computer Science perspective. Different topics of inquiry were distinguished and a taxonomy was proposed to organize them. For each area, it was defined the scope of the works included in it, some of the most representative works, highlighting the discoveries, discussions and challenges of each field.

As seen in the previous sections, computational research in OSN analysis is wide and diverse, enabling the application of techniques from many fields like graph theory, complex networks, dynamic systems, computational simulation, machine learning, natural language processing, data mining, spatio-temporal modeling, among others.

Although many aspects of the presented areas are still being developed, some general movements on the research's course could be identified. The simple characterization of OSN structures, much valued on the first studies, was progressively replaced by studies of users' behaviour on the network and the complex dynamic produced by them. Works using social data for different purposes are also very common, with the knowledge extracted being often considered as a valuable representative of human behaviour or opinion.

Future perspectives

Predicting the next steps of research on OSN is a challenging and risky task. It is even temerarious to predict if the interest on this topic will still be increasing in years to come.

51. `http://www.pinterest.com`
52. `https://www.tumblr.com`
53. `http://www.lastfm.com`

Nonetheless, we will list in the following paragraphs some possibilities of new studies that we believe are worth being explored.

Despite the existance of few works combining information from many social networks, we can notice an increase in the number of theoretical and experimental studies dealing with heterogeneous relationships (e.g.: following, friendship, transportation sharing) from one or more concurrent sources (Gomez-Gardenes et al., 2012; Gómez et al., 2013; Mucha et al., 2010; Sun and Han, 2012). This kind of analysis opens several new roads for research, making possible to have a more complete overview of how individuals interact and influence each other, to better track the evolution of a piece of information and to evaluate how specialized may be the use of different social networks—what may help us to estimate how representative is user behavior in OSNs—, just to cite a few examples. An important aspect to single out is that this data may also be obtained from sources other than OSNs, like surveys and interviews or sensors (e.g.: GPS in smartphones). In fact, we believe that, with the emergence of the "Internet of things", this "offline" data will acquire prominence in computational studies about human behaviour.

In addition to the use of different sources for context awareness, the deeper understanding of how networks evolve during time is also a likely subject to appear in the future. Most studies still consider the network structure as a fixed object, ignoring its transformation and plasticity. The limitation of current methods may be seen in information diffusion analysis, for instance, as the disregard of when a connection is active may create paths that are not temporally consistent and reduce artificially the distance between individuals. More work is required to understand what are the transformations that take place on each kind of network, their impact on the processes observed in complex systems and how such processes influence the evolution of networks themselves.

The knowledge drawn from online social networks may impact not only Computer Science, but it may provoke a revolution in Social Sciences. The burgeoning cross-disciplinary field of Computational Social Science benefits from computational methods, as multi-agent based models, network analysis and machine learning, in order to build a fast, data-driven science. The program of this new data intensive discipline intends to make use of partially structured data available in the Internet, in order to validate and complement existing social theories, or even to propose new research explanations to social phenomena. The use of data from OSNs can not only make much faster the currently time-consuming process of gathering social data, but it may also improve the reproducibility of research in Social Sciences, as every step of the research—from data collection to its analysis—may be audited and reproduced by external agents.

Challenges

Even though the volume of work analyzing OSNs is significant, the area still presents some open challenges, that deserve to be further addressed by researchers.

One initial challenge is associated with the tools and methodologies used. We see that most approaches of OSN studies (specially social data analysis) focus on characteristics of users or messages, but few have a more systemic view, approaching network effects. Therefore, we believe that there is a promising niche to be further explored using methods from Complex Systems and Network Science, trying to understand, for instance, the roles of topology, homophily, heterogeneity in individual behaviors and collective cognition in

such social systems. This kind of research, however, demands tools and strategies yet to be discovered and experienced. More effort, thus, is required to build a robust theoretical framework to tackle those problems adequately.

After approximately ten years in the spotlight, OSNs are still a topic of interest of general media and academia. Buzzwords like "social", "big data" and "complexity" are increasingly popular and the amount of new scientific papers related to them grows each year. At the same time that more discoveries are made it gets more difficult to properly select relevant works and validate new results presented in literature. One of the main aims of this work is, precisely, to help researchers with the task of organizing and selecting material on OSN analysis.

The lack of ethical considerations in most of the observed works is left as our final remark. Even though we focused on computational approaches to online social networks, the information collected and the knowledge produced by the works we analysed have direct implications on societies. For example, the theories and methods developed in this research area can, potentially, be used in harmful ways by authoritarian regimes or abusive advertising campaigns. Privacy is also an important issue as, by analysing public data and behaviours in OSNs, data scientists may uncover implicit information about specific individuals, information that such individuals may have never intended to made public. As OSN analysis is a strongly interdisciplinary field, we believe that this is a current challenge, indispensable to be considered.

Acknowledgements

The authors sincerely thank Romis Attux, Leonardo Maia and Fabrício Olivetti de França for their kind effort of revising this work and contributing with corrections and new insights.

Part of the results presented in this work were obtained through the project "Training in Information Technology", funded by Samsung Eletronics of Amazonia LTDA., using resources from Law of Informatics (Brazilian Federal Law Number 8.248/91).

References

Lada Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3): 211–230, 2003. ISSN 03788733. doi: 10.1016/S0378-8733(03)00009-1.

Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3): 187–203, 2005. ISSN 03788733. doi: 10.1016/j.socnet.2005.01.007.

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, page 835, New York, New York, USA, 2007. ACM Press. ISBN 9781595936547. doi: 10.1145/1242572.1242685.

Harsh Ajmera. Latest social media users stats, facts and numbers for 2014, 2014. URL `http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html`.

Nadeem Akhtar, Hira Javed, and Geetanjali Sengar. Analysis of facebook social network. In *Proceedings - 5th International Conference on Computational Intelligence and Communication Networks, CICN 2013*, pages 451–454. IEEE, 2013. ISBN 9780768550695. doi: 10.1109/CICN.2013.99.

Ofer Arazy, Nanda Kumar, and Bracha Shapira. Improving social recommender systems. *IT Professional*, 11(4):38–44, 2009. ISSN 1520-9202. doi: 10.1109/MITP.2009.76.

Sebastien Ardon, Amitabha Bagchi, Anirban Mahanti, Amit Ruhela, Aaditeshwar Seth, Rudra Mohan Tripathy, and Sipat Triukose. Spatio-temporal and events based analysis of topic popularity in twitter. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, pages 219–228, New York, New York, USA, 2013. ACM Press. ISBN 9781450322638. doi: 10.1145/2505515.2505525.

Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499. IEEE, 2010. ISBN 978-1-4244-8482-9. doi: 10.1109/WI-IAT.2010.63.

Sitaram Asur, Louis Yu, and Bernardo a. Huberman. What trends in chinese social media. *SSRN Electronic Journal*, 2011. ISSN 1556-5068. doi: 10.2139/ssrn.1888779.

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 44, New York, New York, USA, 2006. ACM Press. ISBN 1595933395. doi: 10.1145/1150402.1150412.

Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and curating conversation threads. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, page 13, New York, New York, USA, 2013. ACM Press. ISBN 9781450318693. doi: 10.1145/2433396.2433401.

Eytan Bakshy, Brian Karrer, and Lada Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the tenth ACM conference on Electronic commerce - EC '09*, page 325, New York, New York, USA, 2009. ACM Press. ISBN 9781605584584. doi: 10.1145/1566374.1566421.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, page 65, New York, New York, USA, 2011. ACM Press. ISBN 9781450304931. doi: 10.1145/1935826.1935845.

Peng Bao, Hua-Wei Shen, Junming Huang, and Xueqi Cheng. Popularity prediction in microblogging network: A case study on sina weibo. *arXiv preprint arXiv:1304.4324*, pages 2–3, 2013.

Albert-Laszlo Barabási. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999. ISSN 00368075. doi: 10.1126/science.286.5439.509.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 1–17, 2011. doi: 10.1.1.221.2822.

Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference - IMC '09*, page 49, New York, New York, USA, 2009. ACM Press. ISBN 9781605587714. doi: 10.1145/1644893.1644900.

Michael S. Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, page 21, New York, New York, USA, 2013. ACM Press. ISBN 9781450318990. doi: 10.1145/2470654.2470658.

Alessandro Bessi, Antonio Scala, Luca Rossi, Qian Zhang, and Walter Quattrociocchi. The economy of attention in the age of (mis)information. *Journal of Trust Management*, 1 (1):12, 2014. ISSN 2196-064X. doi: 10.1186/s40493-014-0012-y.

Johan Bollen, Bruno Goncalves, Guangchen Ruan, and Huina Mao. Happiness is assortative in online social networks. *Artificial life*, 17(3):237–251, 2011a. ISSN 1064-5462. doi: 10.1162/artl_a_00034.

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011b. ISSN 18777503. doi: 10.1016/j.jocs.2010.12.007.

Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz, Francisco Sanz, Fermín Serrano, Cristina Viñas, Alfonso Tarancón, and Yamir Moreno. Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. *PloS one*, 6(8):e23883, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0023883.

Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. Locating privileged spreaders on an online social network. *Physical Review E*, 85(6):066123, 2012. ISSN 1539-3755. doi: 10.1103/PhysRevE.85.066123.

Javier Borge-Holthoefer, Raquel a Baños, Sandra González-Bailón, and Yamir Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, 2013. ISSN 2051-1310. doi: 10.1093/comnet/cnt006.

Danah Boyd. Big data: Opportunities for computational and social sciences. http://www.zephoria.org/thoughts/archives/2010/04/17/big-data-opportunities-for-computational-and-social-sciences.html, 2010. URL `http://www.zephoria.org/thoughts/archives/2010/04/17/big-data-opportunities-for-computational-and-social-sciences.html`.

Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010. ISBN 978-1-4244-5509-6. doi: 10.1109/HICSS.2010.412.

John Bragin. *Complexity: A Guided Tour*, volume 13. Oxford University Press, New York, NY, USA, 2010. ISBN 9780195124415. doi: 10.1063/1.3326990.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web - WWW '11*, page 675, New York, New York, USA, 2011. ACM, ACM Press. ISBN 9781450306324. doi: 10.1145/1963405.1963500.

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 - MDMKDD '10*, pages 1–10, New York, New York, USA, 2010. ACM Press. ISBN 9781450302203. doi: 10.1145/1814245.1814249.

Damon Centola. The spread of behavior in an online social network experiment. *Science (New York, N.Y.)*, 329(5996):1194–7, 2010. ISSN 1095-9203. doi: 10.1126/science.1185231.

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007. ISSN 0002-9602. doi: 10.1086/521848.

Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web - WWW '09*, page 721, New York, New York, USA, 2009. ACM Press. ISBN 9781605584874. doi: 10.1145/1526709.1526806.

Meeyoung Cha, Hamed Haddai, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter : The million follower fallacy. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 10:10–17, 2010. doi: 10.1.1.167.192.

Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, page 1185, New York, New York, USA, 2010a. ACM Press. ISBN 9781605589299. doi: 10.1145/1753326.1753503.

Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '10*, pages 1185–1194, New York, New York, USA, 2010b. ACM Press. ISBN 9781605589299. doi: 10.1145/1753326.1753503.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 759, New York, New York, USA, 2010. ACM Press. ISBN 9781450300995. doi: 10.1145/1871437.1871535.

Marc Cheong and Sid Ray. A literature review of recent microblogging developments. *Victoria, Australia: Clayton School of Information Technology, Monash University.*, 2011.

Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. Comparison of online social relations in volume vs interaction. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement conference - IMC '08*, page 57, New York, New York, USA, 2008. ACM Press. ISBN 9781605583341. doi: 10.1145/1452520.1452528.

Claudio Cioffi-Revilla. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):259–271, 2010. ISSN 19395108. doi: 10.1002/wics.95.

David Combe, Christine Largeron, Elod Egyed-Zsigmond, and Mathias Géry. A comparative study of social network analysis tools. In *International Workshop on Web Intelligence and Virtual Enterprises*, volume 2, page 1, 2010.

R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J. P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel, and D. Helbing. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346, 2012. ISSN 1951-6355. doi: 10.1140/epjst/e2012-01697-8.

Luciano F. Costa, Osvaldo N. Oliveira, Gonzalo Travieso, Francisco A. Rodrigues, Paulino R. Villas Boas, Lucas Antiqueira, Matheus P. Viana, and Luis E. C. da Rocha. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Advances in Physics*, 60(3):103, 2007. ISSN 0001-8732. doi: 10.1080/00018732.2011.572452.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, page 1695, 2006.

Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 115–122, New York, New York, USA, 2010a. ACM Press. ISBN 9781450302173. doi: 10.1145/1964858.1964874.

Aron Culotta. Detecting influenza outbreaks by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 115–122, New York, New York, USA, 2010b. ACM Press. ISBN 9781450302173. doi: 10.1145/1964858.1964874.

Ithiel de Sola Pool and Manfred Kochen. Contacts and influence. *Social Networks*, 1(1):5–51, 1978. ISSN 03788733. doi: 10.1016/0378-8733(78)90011-4.

Martin Dillon. Introduction to modern information retrieval. *Information Processing & Management*, 19(6):402–403, 1983. ISSN 03064573. doi: 10.1016/0306-4573(83)90062-6.

Peter Sheridan Dodds, Kameroncker 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. ISSN 1932-6203. doi: 10.1371/journal.pone.0026752.

P Alex Dow and Adrien Friggeri. The anatomy of large facebook cascades. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 145–154, 2013.

R.I.M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992. ISSN 00472484. doi: 10.1016/0047-2484(92)90081-J.

Facebook. Company info — facebook newsroom, 2014. URL `http://newsroom.fb.com/company-info/`.

Adrien Friggeri, La Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014.

Qi Gao, Fabian Abel, Geert Jan Houben, and Yong Yu. A comparative study of users' microblogging behavior on sina weibo and twitter. In Judith Masthoff, Bamshad Mobasher, Michel C. Desmarais, and Roger Nkambou, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7379 LNCS, pages 88–101, Berlin, Heidelberg, 2012. Springer. ISBN 9783642314537. doi: 10.1007/978-3-642-31454-4_8.

David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. Social resilience in online communities. In *Proceedings of the first ACM conference on Online social networks - COSN '13*, volume 40, pages 39–50, New York, New York, USA, 2013. ACM Press. ISBN 9781450320849. doi: 10.1145/2512938.2512946.

Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, 31(6):649–679, 2013. ISSN 0894-4393, 1552-8286. doi: 10.1177/0894439313493979.

M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266–6282, 2013. ISSN 09574174. doi: 10.1016/j.eswa.2013.05.057.

Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–6, 2002. ISSN 0027-8424. doi: 10.1073/pnas.122653799.

Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce - EC '12*, volume 1, page 623, New York, New York, USA, 2012. ACM Press. ISBN 9781450314152. doi: 10.1145/2229012.2229058.

Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan Watts. The structural virality of online diffusion. *Preprint*, 2013.

Jennifer Golbeck and Derek Hansen. A method for computing political preference among twitter followers. *Social Networks*, 36:177–184, 2014. ISSN 03788733. doi: 10.1016/j.socnet.2013.07.004.

Benjamin Golub and Matthew O Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24):10833–6, 2010. ISSN 1091-6490. doi: 10.1073/pnas.1000814107.

S. Gómez, A. Díaz-Guilera, J. Gómez-Gardeñes, C. J. Pérez-Vicente, Y. Moreno, and A. Arenas. Diffusion dynamics on multiplex networks. *Physical Review Letters*, 110 (2):028701, 2013. ISSN 0031-9007. doi: 10.1103/PhysRevLett.110.028701.

Jesus Gomez-Gardenes, Irene Reinares, Alex Arenas, and Luis Mario Floría. Evolution of cooperation in multiplex networks. *Scientific reports*, 2:620, 2012. ISSN 2045-2322. doi: 10.1038/srep00620.

Manuel Gomez-Rodriguez, Jure Leskovec, Andreas Krause, Manuel Gomez Rodriguez, Jure Leskovec, Andreas Krause, Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data*, 5(4):1–37, 2012. ISSN 15564681. doi: 10.1145/2086737.2086741.

Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: validation of dunbar's number. *PloS one*, 6(8):e22656, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0022656.

S. Gonzalez-Bailon, Javier Borge-Holthoefer, and Yamir Moreno. Broadcasters and hidden influentials in online protest diffusion. *American Behavioral Scientist*, 57(7):943–965, 2013. ISSN 0002-7642. doi: 10.1177/0002764213479371.

Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1: 197, 2011. ISSN 2045-2322. doi: 10.1038/srep00197.

Przemyslaw a. Grabowicz, José J. Ramasco, Esteban Moro, Josep M. Pujol, and Victor M. Eguiluz. Social features of online networks: the strength of intermediary ties in online social media. *PloS one*, 7(1):e29358, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0029358.

D. Gruhl, David Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, 6(2):43–52, 2004. ISSN 19310145. doi: 10.1145/1046456.1046462.

Daniel Gruhl, R Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, page 78, New York, New York, USA, 2005. ACM Press. ISBN 159593135X. doi: 10.1145/1081870.1081883.

Marco Guerini, Jacopo Staiano, and Davide Albanese. Exploring image virality in google plus. In *2013 International Conference on Social Computing*, pages 671–678. IEEE, 2013. ISBN 978-0-7695-5137-1. doi: 10.1109/SocialCom.2013.101.

Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel a. Zighed. Information diffusion in online social networks. *ACM SIGMOD Record*, 42(1):17, 2013. ISSN 01635808. doi: 10.1145/2503792.2503797.

Zhengbiao Guo, Zhitang Li, and Hao Tu. Sina microblog: An information-driven online social network. In *2011 International Conference on Cyberworlds*, pages 160–167. IEEE, 2011. ISBN 978-1-4577-1453-5. doi: 10.1109/CW.2011.12.

Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In *SIAM International Conference on Data Mining*, pages 153–164. Citeseer, 2012.

Aric A Hagberg, Daniel A Schult, and Pieter J Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, 2008.

Benjamin Herd, Simon Miles, Peter Mcburney, and Michael Luck. *Multi-Agent-Based Simulation XIV*, volume 8235. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-54782-9. doi: 10.1007/978-3-642-54783-6.

Tuan-anh Hoang and Ee-peng Lim. Virality and susceptibility in information diffusions. *Artificial Intelligence*, pages 146–153, 2012.

Tad Hogg and Kristina Lerman. Stochastic models of user-contributory web sites. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 50–57, 2009a. doi: 10.1145/1772690.1772754.

Tad Hogg and Kristina Lerman. Stochastic models of user-contributory web sites. *Proceedings of the 19th international conference on World wide web - WWW '10*, pages 50–57, 2009b. doi: 10.1145/1772690.1772754.

Courtenay Honeycutt and Susan C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2009. ISBN 978-0-7695-3450-3. doi: 10.1109/HICSS.2009.89.

Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web - WWW '11*, page 57, New York, New York, USA, 2011. ACM Press. ISBN 9781450306379. doi: 10.1145/1963192.1963222.

Daniel J. Howard and Charles Gengler. Emotional contagion effects on product attitudes. *Journal of Consumer Research*, 28(2):189–201, 2001. ISSN 0093-5301. doi: 10.1086/322897.

Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu K L Ma. Breaking news on twitter. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 2751, New York, New York, USA, 2012. ACM, ACM Press. ISBN 9781450310154. doi: 10.1145/2207676.2208672.

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *International Conference on World Wide Web*, pages 607–617, Rio de Janeiro, Brazil, 2013a. International World Wide Web Conferences Steering Committee. ISBN 9781450320351.

Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, volume 1, page 537, New York, New York, USA, 2013b. ACM Press. ISBN 9781450318693. doi: 10.1145/2433396.2433465.

Bernardo a. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *SSRN Electronic Journal*, 14, 2008. ISSN 1556-5068. doi: 10.2139/ssrn.1313405.

Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188, 2009. ISSN 15322882. doi: 10.1002/asi.21149.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07*, pages 56–65, New York, New York, USA, 2007. ACM Press. ISBN 9781595938480. doi: 10.1145/1348549.1348556.

Adam D I Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24):8788–90, 2014. ISSN 1091-6490. doi: 10.1073/pnas.1320040111.

Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks - WOSP '08*, page 19, New York, New York, USA, 2008. ACM Press. ISBN 9781605581828. doi: 10.1145/1397735.1397741.

Tobias Kuhn, Matjaž Perc, and Dirk Helbing. Inheritance patterns in citation networks reveal scientific memes. *Physical Review X*, 4(4):041036, 2014. ISSN 2160-3308. doi: 10.1103/PhysRevX.4.041036.

Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, page 553, New York, New York, USA, 2010a. ACM Press. ISBN 9781450300551. doi: 10.1145/1835804.1835875.

Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In Philip S Yu, Jiawei Han, and Christos Faloutsos, editors, *Link Mining: Models, Algorithms, and Applications*, pages 337–357. Springer New York, New York, NY, 2010b. ISBN 978-1-4419-6514-1, 978-1-4419-6515-8. doi: 10.1007/978-1-4419-6515-8.

Sanjeev Kumar. Analyzing the facebook workload. In *2012 IEEE International Symposium on Workload Characterization (IISWC)*, pages 111–112. IEEE, 2012. ISBN 978-1-4673-4532-3. doi: 10.1109/IISWC.2012.6402911.

Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, page 2335, New York, New York, USA, 2012. ACM Press. ISBN 9781450311564. doi: 10.1145/2396761.2398634.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 591, New York, New York, USA, 2010. ACM Press. ISBN 9781605587998. doi: 10.1145/1772690.1772751.

Haewoon Kwak, Hyunwoo Chun, and Sue Moon. Fragile online relationship. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 1091, New York, New York, USA, 2011. ACM Press. ISBN 9781450302289. doi: 10.1145/1978942.1979104.

Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–22, 2012. ISSN 21576904. doi: 10.1145/2337542.2337557.

Thomas Lansdall-Welfare, Vasileios Lampos, and Nello Cristianini. Effects of the recession on public mood in the uk. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 1221, New York, New York, USA, 2012. ACM Press. ISBN 9781450312301. doi: 10.1145/2187980.2188264.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Social science. computational social science. *Science (New York, N.Y.)*, 323(5915):721–3, 2009. ISSN 1095-9203. doi: 10.1126/science.1167742.

Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. Who will retweet this ? automatically identifying and engaging strangers on twitter to spread information. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, pages 247—-256, New York, New York, USA, 2014. ACM Press. ISBN 9781450321846. doi: 10.1145/2557500.2557502.

Kristina Lerman and Rumi Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 90–97, 2010. ISSN 00846570. doi: 10.1146/annurev.an.03.100174.001431.

Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, page 915, New York, New York, USA, 2008. ACM Press. ISBN 9781605580852. doi: 10.1145/1367497.1367620.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data*

mining - KDD '05, page 177, New York, New York, USA, 2005. ACM Press. ISBN 159593135X. doi: 10.1145/1081870.1081893.

Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. *SIAM International Conference on Data Mining (SDM)*, 2007. ISSN 0038-0644. doi: 10.1.1.103.8339.

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 - KDD '09*, volume 1, pages 497–506, New York, New York, USA, 2009. ACM Press. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557077.

David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12):4633–8, 2008. ISSN 1091-6490. doi: 10.1073/pnas.0708471105.

Fengkun Liu and Hong Joo Lee. Use of social network information to enhance collaborative filtering performance. *Expert Systems with Applications*, 37(7):4772–4778, 2010. ISSN 09574174. doi: 10.1016/j.eswa.2009.12.061.

Sophia B Liu. The rise of curated crisis content. *Iscram*, pages 1–6, 2010.

Rong Lu and Qing Yang. Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, 2:327–332, 2012. ISSN 20103700. doi: 10.7763/IJMLC.2012.V2.139.

Hao Ma, Tom Chao Zhou, Michael R. Lyu, and Irwin King. Improving recommender systems by incorporating social contextual information. *ACM Transactions on Information Systems*, 29(2):1–23, 2011. ISSN 10468188. doi: 10.1145/1961209.1961212.

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. ISSN 15322882. doi: 10.1002/asi.22844.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001. ISSN 0360-0572. doi: 10.1146/annurev.soc.27.1.415.

Nasrullah Memon and Reda Alhajj. *From sociology to computing in social networks: Theory, foundations and applications*. Springer Vienna, Vienna, 2010. ISBN 9783709102930. doi: 10.1007/978-3-7091-0294-7.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 71–79, New York, New York, USA, 2010. ACM Press. ISBN 9781450302173. doi: 10.1145/1964858.1964869.

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the*

7th ACM SIGCOMM conference on Internet measurement - IMC '07, page 29, New York, New York, USA, 2007. ACM Press. ISBN 9781595939081. doi: 10.1145/1298306.1298311.

CRAIG MORGAN. In this issue. *Psychological Medicine*, 39(12):1933, 2009. ISSN 0033-2917. doi: 10.1017/S0033291709991759.

Peter J Mucha, Thomas Richardson, Kevin Macon, Mason a Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science (New York, N.Y.)*, 328(5980):876–8, 2010. ISSN 1095-9203. doi: 10.1126/science.1184819.

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 - KDD '12*, page 33, New York, New York, USA, 2012. ACM Press. ISBN 9781450314626. doi: 10.1145/2339530.2339540.

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast. In *Proceedings of the 3rd International Web Science Conference on - WebSci '11*, pages 1–7, New York, New York, USA, 2011. ACM Press. ISBN 9781450308557. doi: 10.1145/2527031.2527052.

Graham Neubig, Y Matsubayashi, M Hagiwara, and K Murakami. Safety information mining-what can nlp do in a disaster. In *IJCNLP*, pages 965–973, 2011.

G.J. Nichols and J.A. Fisher. Processes, facies and architecture of fluvial distributary system deposits. *Sedimentary Geology*, 195(1-2):75–90, 2007. ISSN 00370738. doi: 10.1016/j.sedgeo.2006.07.004.

Josh Ong. China's sina weibo grew 73% in 2012 to 500m accounts, 2013. URL `http://thenextweb.com/asia/2013/02/21/chinas-sina-weibo-grew-73-in-2012-passing-500-million-registered-accounts/`.

L Page, S Brin, R Motwani, and T Winograd. The pagerank citation ranking:bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, 1998.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Lrec*, volume Proceeding, pages 1320–1326, Valletta, Malta, 2010. European Languages Resources Association (ELRA). ISBN 2951740867. doi: 10.1371/journal.pone.0026624.

Tiago P Peixoto. The graph-tool python library. *figshare*, 2014. doi: 10.6084/m9.figshare.1164194.

S Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 13:586–589, 2011.

Fotis Psallidas, Luis Gravano, and Cornell Tech. Effective event identification in social media. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 36(3):42–50, 2013. doi: 10.1007/BF00183540.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it : Identifying misinformation in microblogs. In *Proceeding of the 2011 Conference on Empirical Methods in Natural Language Processing - 'EMNLP*, pages 1589–1599, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics, Association for Computational Linguistics. ISBN 978-1-937284-11-4.

Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. Microblogging after a major disaster in china. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, page 25, New York, New York, USA, 2011. ACM Press. ISBN 9781450305563. doi: 10.1145/1958824.1958830.

Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268, 2012. ISSN 1574-020X. doi: 10.1007/s10579-012-9196-x.

Stuart a Rice. The identification of blocs in small political bodies. *The American Political Science Review*, 21(3):619, 1927. ISSN 00030554. doi: 10.2307/1945514.

Richard Rogers. Debanalizing twitter. In *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, pages 356–365, New York, New York, USA, 2013. ACM Press. ISBN 9781450318891. doi: 10.1145/2464464.2464511.

Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics. In *Proceedings of the 20th international conference on World wide web - WWW '11*, page 695, New York, New York, USA, 2011a. ACM Press. ISBN 9781450306324. doi: 10.1145/1963405.1963503.

Daniel M. Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. *arXiv preprint arXiv:1112.1115*, page 11, 2011b.

Rintaro Saito, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker. A travel guide to cytoscape plugins. *Nature methods*, 9(11):1069–1076, 2012.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users. In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 851, New York, New York, USA, 2010. ACM Press. ISBN 9781605587998. doi: 10.1145/1772690.1772777.

Marcel Salathé, Duy Q Vu, Shashank Khandelwal, and David R. Hunter. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science*, 2(1):4, 2013. ISSN 2193-1127. doi: 10.1140/epjds16.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science (New York, N.Y.)*, 311(5762):854–6, 2006. ISSN 1095-9203. doi: 10.1126/science.1121066.

Aleksandra Sarcevic, Leysia Palen, Joanne White, Kate Starbird, Mossaab Bagdouri, and Kenneth Anderson. "beacons of hope" in decentralized coordination. In *Proceedings*

of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12, page 47, New York, New York, USA, 2012. ACM, ACM Press. ISBN 9781450310864. doi: 10.1145/2145204.2145217.

Kazutoshi Sasahara, Yoshito Hirata, Masashi Toyoda, Masaru Kitsuregawa, and Kazuyuki Aihara. Quantifying collective attention from tweet stream. *PLoS ONE*, 8(4):e61823, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0061823.

Devavrat Shah and Tauhid Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, 2011. ISSN 00189448. doi: 10.1109/TIT.2011.2158885.

Paulo Shakarian, Sean Eyre, and Damon Paulo. A scalable heuristic for viral marketing under the tipping model. *arXiv preprint arXiv:1309.2963*, 3(4):37, 2013. ISSN 1869-5450. doi: 10.1007/s13278-013-0135-7.

David a. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. Peaks and persistence. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, pages 355–358, New York, New York, USA, 2011. ACM Press. ISBN 9781450305563. doi: 10.1145/1958824.1958878.

Marc a Smith, Lee Rainie, Itai Himelboim, and Ben Shneiderman. Mapping twitter topic networks: From polarized crowds to community clusters. *The Pew Research Center*, pages 1–57, 2014.

Kate Starbird and L Palen. *Pass it on?: Retweeting in mass emergency*. International Community on Information Systems for Crisis Response and Management, 2010. doi: 10.1111/j.1556-4029.2009.01231.x.

Kate Starbird and Leysia Palen. "voluntweeters". In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 1071, New York, New York, USA, 2011. ACM Press. ISBN 9781450302289. doi: 10.1145/1978942.1979102.

Kate Starbird and Leysia Palen. (how) will the revolution be retweeted? In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, page 7, New York, New York, USA, 2012. ACM, ACM Press. ISBN 9781450310864. doi: 10.1145/2145204.2145212.

Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. Chatter on the red. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*, page 241, New York, New York, USA, 2010. ACM Press. ISBN 9781605587950. doi: 10.1145/1718918.1718965.

Christian L Staudt, Aleksejs Sazonovs, and Henning Meyerhenke. Networkit: An interactive tool suite for high-performance network analysis. *arXiv preprint arXiv:1403.3005*, 2014.

Stefan Stieglitz and Linh Dang-Xuan. Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Sciences*, pages 3500–3509. IEEE, 2012. ISBN 978-1-4577-1925-7. doi: 10.1109/HICSS.2012.476.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE, 2010. ISBN 978-1-4244-8439-3. doi: 10.1109/SocialCom.2010.33.

Eric Sun, Itamar Rosenn, Cameron a Marlow, and Thomas M Lento. Gesundheit ! modeling contagion through facebook news feed mechanics of facebook page diffusion. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 146–153, 2009. ISBN 978-1-57735-421-5.

Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: Principles and methodologies. *Morgan & Claypool Publishers*, 3(2):1–159, 2012. ISSN 2151-0067. doi: 10.2200/S00433ED1V01Y201207DMK005.

Gabor Szabo and Bernardo a. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80, 2010. ISSN 00010782. doi: 10.1145/1787234.1787254.

Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi. Discovering emerging topics in social streams via link-anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):120–130, 2014. ISSN 1041-4347. doi: 10.1109/TKDE.2012.239.

Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of twitter networks. *Social Networks*, 34(1):73–81, 2012. ISSN 03788733. doi: 10.1016/j.socnet.2011.05.006.

John Tang, Mirco Musolesi, Cecilia Mascolo, and Vito Latora. Characterising temporal distance and reachability in mobile and online social networks. *ACM SIGCOMM Computer Communication Review*, 40(1):118, 2010. ISSN 01464833. doi: 10.1145/1672308.1672329.

Rudra M. Tripathy, Amitabha Bagchi, and Sameep Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 1817, New York, New York, USA, 2010. ACM, ACM Press. ISBN 9781450300995. doi: 10.1145/1871437.1871737.

Oren Tsur and Ari Rappoport. What's in a hashtag? In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, page 643, New York, New York, USA, 2012. ACM Press. ISBN 9781450307475. doi: 10.1145/2124295.2124320.

Andranik Tumasjan, To Sprenger, Pg Sandner, and Im Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 10:178–185, 2010. ISSN 00219258.

Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo a. Huberman. E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, 21(2):143–153, 2005. ISSN 0197-2243. doi: 10.1080/01972240590925348.

Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events. In *Proceedings of the 28th international conference on Human*

factors in computing systems - CHI '10, page 1079, New York, New York, USA, 2010. ACM Press. ISBN 9781605589299. doi: 10.1145/1753326.1753486.

Frank Edward Walter, Stefano Battiston, and Frank Schweitzer. A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems*, 16(1):57–74, 2007. ISSN 1387-2532. doi: 10.1007/s10458-007-9021-x.

D J Watts and S H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684):440–2, 1998. ISSN 0028-0836. doi: 10.1038/30918.

Duncan J. Watts. Everything is obvious: How common sense fails us. *Random House LLC*, page 368, 2012.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, page 261, New York, New York, USA, 2010. ACM Press. ISBN 9781605588896. doi: 10.1145/1718487.1718520.

Jianshu Weng, Yuxia Yao, Erwin Leonardi, Francis Lee, and Bu-sung Lee. Event detection in twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2:335, 2012. ISSN 2045-2322. doi: 10.1038/srep00335.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3:2522, 2013. ISSN 2045-2322. doi: 10.1038/srep02522.

Wikipedia. Social network analysis software - wikipedia, the free encyclopedia, 2015. URL `http://en.wikipedia.org/w/index.php?title=Social_network_analysis_software&oldid=651703528`.

Dennis M Wilkinson. Strong regularities in online peer production. In *Proceedings of the 9th ACM conference on Electronic commerce - EC '08*, page 302, New York, New York, USA, 2008. ACM Press. ISBN 9781605581699. doi: 10.1145/1386790.1386837.

Shirley a. Williams, Melissa M. Terras, and Claire Warwick. What do people study when they study twitter? classifying twitter related academic papers. *Journal of Documentation*, 69(3):384–410, 2013. ISSN 0022-0418. doi: 10.1108/JD-03-2012-0027.

Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *Proceedings of the fourth ACM european conference on Computer systems - EuroSys '09*, page 205, New York, New York, USA, 2009. ACM Press. ISBN 9781605584829. doi: 10.1145/1519065.1519089.

Felix Ming Fai Wong, Soumya Sen, and Mung Chiang. Why watching movie tweets won't tell the whole story? In *Proceedings of the 2012 ACM workshop on Workshop on online*

social networks - WOSN '12, page 61, New York, New York, USA, 2012. ACM Press. ISBN 9781450314800. doi: 10.1145/2342549.2342564.

Fang Wu and Bernardo a Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45):17599–601, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0704916104.

Shaomei Wu, Jake M. Hofman, Winter a. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web - WWW '11*, page 705, New York, New York, USA, 2011. ACM Press. ISBN 9781450306324. doi: 10.1145/1963405.1963504.

Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web - WWW '10*, volume 55, page 981, New York, New York, USA, 2010. ACM Press. ISBN 9781605587998. doi: 10.1145/1772690.1772790.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12*, volume 2, pages 1–7, New York, New York, USA, 2012. ACM Press. ISBN 9781450315463. doi: 10.1145/2350190.2350203.

Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608, Sydney, NSW, 2010. IEEE. ISBN 978-1-4244-9131-5. doi: 10.1109/ICDM.2010.22.

Xiwang Yang, Yang Guo, and Yong Liu. Bayesian-inference-based recommendation in online social networks. *IEEE Transactions on Parallel and Distributed Systems*, 24(4):642–651, 2013. ISSN 1045-9219. doi: 10.1109/TPDS.2012.192.

Xue Zhang, Hauke Fuehres, and Peter a. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26: 55–62, 2011. ISSN 18770428. doi: 10.1016/j.sbspro.2011.10.562.

Changtao Zhong, Sunil Shah, and Nishanth Sastry. Sharing the loves : Understanding the how and why of online content curation. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media Sharing*, pages 659–667, 2013.

Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, and Vwani Roychowdhury. Information resonance on twitter. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 123–131, New York, New York, USA, 2010. ACM, ACM Press. ISBN 9781450302173. doi: 10.1145/1964858.1964875.