David Burth Kurka

Online Social Networks: knowledge extraction from information diffusion and analysis of spatio-temporal phenomena

Redes Sociais Online: extração de conhecimento e análise espaço-temporal de eventos de difusão de informação.

Campinas

2015

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

David Burth Kurka

Online Social Networks: knowledge extraction from information diffusion and analysis of spatio-temporal phenomena

Redes Sociais Online: extração de conhecimento e análise espaço-temporal de eventos de difusão de informação.

Master dissertation presented to the School of Electrical and Computer Engineering of the University of Campinas to obtain the M.Sc. degree in Electrical Engineering, in the area of Computer Engineering.
Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Supervisor: Prof. Dr. Fernando José Von Zuben

Este exemplar corresponde à versão final da tese defendida pelo aluno David Burth Kurka, e orientada pelo Prof. Dr. Fernando José Von Zuben

---

Campinas
2015

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

K965o

Kurka, David Burth, 1988-
Online social networks : knowledge extraction from information diffusion and analysis of spatio-temporal phenomena / David Burth Kurka. – Campinas, SP : [s.n.], 2015.

Orientador: Fernando José Von Zuben.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Redes sociais on-line. 2. Sistemas complexos. 3. Aprendizado de máquina. I. Von Zuben, Fernando José,1968-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Redes sociais online : extração de conhecimento e análise espaço-temporal de eventos de difusão de informação
Palavras-chave em inglês:
Online social networks
Complex systems
Machine learning
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Fernando José Von Zuben [Orientador]
Fabrício Olivetti de França
Eduardo Alves do Valle Junior
Data de defesa: 08-05-2015
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: David Burth Kurka

Data da Defesa: 8 de maio de 2015

Título da Tese: "Online Social Networks: Knowledge Extraction from Information Diffusion and Analysis of Spatio-Temporal Phenomena (Redes Sociais Online: Extração de Conhecimento e Análise Espaço-Temporal de Eventos de Difusão de Informação)"

Prof. Dr. Fernando José Von Zuben (Presidente): Fernando José Von Zuben

Prof. Dr. Fabricio Olivetti de França: Fabricio Olivetti de França

Prof. Dr. Eduardo Alves do Valle Junior: Eduardo A. do Valle Jr.

Abstract

With the advent and popularization of Online Social Networks and Social Networking Services, computer science researchers have found fertile field for the development of studies using large volumes of data, multiple agents models, and spatio-temporal dynamics. However, even with a significant amount of published research on the subject, there are still aspects of social networks whose explanation is incipient. In order to deepen the knowledge of the area, this work investigates phenomena of collective sharing on the network, characterizing information diffusion events. From the observation of real data obtained from the online service Twitter, we collect, model and characterize such events, from a perspective of complex systems. Finally, using machine learning and computational data analysis, patterns are found on the network's spatio-temporal processes, making it possible to classify a message's topic from users behaviour and to characterize individual behaviour from social connections.

Keywords: Online Social Networks; Complex Systems; Machine Learning

Resumo

Com o surgimento e a popularização de Redes Sociais Online e de Serviços de Redes Sociais, pesquisadores da área de computação têm encontrado um campo fértil para o desenvolvimento de trabalhos com grande volume de dados, modelos envolvendo múltiplos agentes e dinâmicas espaço-temporais. Entretanto, mesmo com um significativo elenco de pesquisas já publicadas no assunto, ainda existem aspectos das redes sociais cuja explicação é incipiente. Visando o aprofundamento do conhecimento da área, este trabalho investiga fenômenos de compartilhamento coletivo na rede, que caracterizam eventos de difusão de informação. A partir da observação de dados reais oriundos do serviço online Twitter, tais eventos são modelados, caracterizados e analisados sob a perspectiva de sistemas complexos. Com o uso de técnicas de aprendizado de máquina e análise computacional de dados, são encontrados padrões nos processos espaço-temporais da rede, tornando possível a construção de classificadores de mensagens baseados em comportamento e a caracterização de comportamentos individuais, a partir de conexões sociais.

Palavras-chaves: Redes Sociais Online; Sistemas Complexos; Aprendizado de Máquina

Contents

This dissertation is dedicated to my family and wife.

Acknowledgements

Talvez uma das conclusões mais importantes que tiro deste período de investigação do mestrado é o reforço da crença do quanto somos influenciados e dependentes de nossos pares. Por mais que consiga ver esforço pessoal nessa trajetória, não ouso deixar de reconhecer que, se chego a algum lugar novo, é graças a muitas pessoas.

Por ordem cronológica, gostaria de primeiro agradecer a meus pais, Paulo e Anita e a meu irmão Pedro, pela base de toda educação, motivação e sonhos que eu possa ter. Por mais criativo que eu possa querer ser em minha vida, não há como negar a influência e inspiração vindas de vocês, que sempre farão parte de mim. Pelo mesmo motivo, também agradeço à família estendida, principalmente aqui aos meus avós Luiz e Brasil, que sempre acreditaram muito em mim e não pouparam incentivos para a minha formação.

Tão importante quanto a família, quero me lembrar dos amigos que têm me ensinado e me influenciado muito. Aos amigos da ABU, que marcaram a minha graduação e o mestrado, sou muito grato por tudo. Em especial, aos moradores da Casa Douglas, com os quais pude dividir ainda mais: Bacon, Bember, Carla, Jonas, Pedro Ivo, Mauê, Eliezer, Nathan, Micael, Edu, Gabriel, Davi, Glauber, Daniel e Juari. Aos grandes amigos da IPBG, Caio, Pri Gomes, Elmo, Salatiel, Pri Akemi, Rômulo, Sabrina, Tovo, Daniel.

Aos colegas da Unicamp, agradeço a Eltermann, Teo Wey, Gaiowsky, Birocchi e Tolstenko. Já tem sido divertido ver quais caminhos esse grupo de sobrenomes esquisitos tem tomado e aguardo ansioso para descobrir o que será de cada um de vocês no futuro!

Aos amigos da APOGEEU, agradeço também pelo bom tempo que tivemos juntos, durante esse tempo: Luiz, Raul, Edgar, Rafael.

Finalmente, um agradecimento muito especial a todos os colegas do LBiC, os

maiores responsáveis pelo meu crescimento e aprendizado nesse período. Considero todas as conversas (sérias e brincadeiras), os Grudes, as parcerias em disciplinas, os conselhos e toda a convivência, como um dos pontos mais altos desses últimos anos. Agradeço especialmente a parceria e amizade do Saullo, a hospitalidade e sensibilidade do Marcos, o coração generoso e sincero da Rosana, os iogurtes do bandejão do Will, as piadas sem graça do Hamilton, a personalidade apaixonada e intensa do Carlos, a inteligência do André, a cultura do Conrado e a determinação da Thalita. Essa combinação de gênios faz de lá um lugar muito único. Também quero lembrar dos outros colegas próximos ao lab, em especial André Virgílio, Alan Caio e Micael, que também se tornou ultimamente um companheiro nas madrugadas de trabalho!

Ainda na Unicamp, agradeço duas pessoas que contribuíram sobremanera para o desenvolvimento desse trabalho. Alan, que apesar de infelizmente não ser considerado formalmente meu co-orientador, por ainda estar desfrutando do seu período de doutoramento, atuou dessa forma desde o início, de forma muito intensa, desde em conversas iniciais sobre temas de pesquisa, até em revisões e orientações sobre todo o material escrito e produzido por mim. Te agradeço enormemente pelo seu empenho e apoio, que foram essenciais para o cumprimento dessas etapas. Fernando Von Zuben, que me ensinou muito tanto como professor, mas como exemplo de profissional e indivíduo. Sua dedicação ao trabalho e ao apoio de seus orientandos é inspiradora e foi essencial para a minha caminhada nesses últimos dois anos.

Quero também agradecer minha amada esposa e companheira Rebeca. Muito do que foi vivido durante esse processo do mestrado foi compartilhado com ela, como noites mal dormidas, planejamentos arriscados e ambições futuras, mas nada foi o bastante para afastá-la ou diminuir seu apoio. A sua companhia até aqui me dá segurança para confiar nos próximos passos, por mais incertos que sejam, que tomaremos.

Por fim, não posso deixar de agradecer a Deus, o qual entendo ser não só o motivador de todas as coisas, mas também o que dá sentido e propósito a toda atividade. Que eu possa ter o privilégio de continuar aprendendo e sendo influenciado por tantas pessoas queridas, pelo resto da minha vida.

"Go to the ant, O sluggard; consider her ways, and be wise. Without having any chief, officer, or ruler, she prepares her bread in summer and gathers her food in harvest."

(Proverbs 6:6-8)

List of Figures

List of Tables

1 Introduction

1.1 Introduction and Motivation

Online social networks can be interpreted as a by-product of the impressive advancement of information technology, and currently experience worldwide, fast growing adhesion as services like *Twitter* and *Facebook* surpass the number of hundreds of millions of users (HOLT, 2013). The average navigation time in such services is also increasing and many Internet sites already contain some form of integration with online social networks. It is difficult to quantify the influence and impact of such services in life beyond the scope of the Internet, but clear effects are already perceived in realms such as personal relations, cultural standards, way of life, education and politics.

Although social interactions have been the object of scientific investigation for many years already (DE SOLA POOL; KOCHEN, 1978), the vast amount of data generated and stored by online networks make the present moment particularly prone to richer and deeper studies on social relations. There is a great interest in understanding the properties and effects of the spatio-temporal dynamics of communication and social interaction, from the point of view of scientific research and innovation, both in virtual and real environments.

One special and intriguing aspect of such virtual environments is the process in which information is generated, broadcast and processed by network users. It is noticeable that meaningful events are created and organized in such communities, without a centralized control, as a result of an intelligent and collective orchestration (SARCEVIC *et al.*, 2012).

This work proposes an in-depth study of the spatio-temporal phenomena of information diffusion and collective mobilization. An effort will be made to identify, model and characterize such events, in order to build models able to describe their

operation. Thus, an understanding of the mechanisms behind those processes and the factors and entities responsible for determining the nature of observable events will be pursued.

The understanding of such emerging phenomena is of common interest to a vast range of research areas. In telecommunications, for example, multi-user network communications is a theme with many open questions, and an ever-growing demand for novel techniques and knowledge paradigms (GAMAL *et al.*, 2011). The spreading of new beliefs and influence networks are of interest to social psychology, publicity, or even politics (HALU *et al.*, 2013; WATTS; DODDS, 2007). Decentralized and parallel processing, in computer sciences, can also benefit from a better understanding of such social phenomena, that could serve as a theoretical and inspirational background for the development of application tools (GODOY; VON ZUBEN, 2013; MITCHELL, 2009).

Also, the intended contribution of this work may have a great innovative impact in different applications. The automatic verification of rumors that are being spread, for example, could be of interest to people and authorities that are in a position to provide timely answers to new facts. It can also be used to verify the veracity of content published in news sources, or even provide feedback and help in predicting the reaction of individuals to the introduction of a new product, that can be of strategic use to public policies or business operations.

1.2 Main Goal of the Research

From the observation of real processes in online social networks, this work aims to develop mathematical and computer models with a potential to identify, predict and analyse behaviour patterns, using network structure information. The focus is on the investigation of how information disseminated and processed throughout social networks.

1.3 Structure of the Text

This work is organized in two main parts: first, the introduction of the theoretical background to the study of online social networks (Chapters 2 and 3) and then the presentation of practical experiments and empirical investigation (Chapters 4 to 7.

Chapter 2 will describe the existing studies, under a computational perspective, that have online social networks as objective of study. As the amount of work is considerable, a taxonomy is suggested, to classify the diverse branches of research. The present work is then situated and detailed within the proposed taxonomy.

Chapter 3 presents what will be our theoretical framework, to describe and understand social networks: *the complex systems theory*. This theory is fundamental for our work, because it put forward a comprehensive perspective to explain global phenomena present in the network, and also a series of mathematical elements, based on graph theory, that can be used to study the network's structure and its behaviour.

Chapter 4 describes the experiments proposed and the details of the methodology that is used on them. Also, the choice of a specific social networking service, for data extraction, is properly justified.

Chapters 5, 6 and 7 deal with experimental data. First, Chapter 5 describes the database collected, with all its properties and peculiarities.

Subsequently, on Chapter 6, a set of experiments are performed, in order to investigate relationships between the network behaviour and properties of the content being spread on it. A classifier capable of predicting a message's subject from the users behaviour is built.

Chapter 7 proposes another set of experiments, investigating if there is a relationship between social structures and behaviour. To verify the hypothesis, message sharing patterns for different groups of users are analysed.

Finally, Chapter 8 outlines the final remarks and also is devoted to a personal viewpoint of the intended contribution of the research.

2 Online Social Network Analysis

This chapter is intended to be a concise version of the survey in Kurka *et al.* (2015).

One of the most revolutionary aspects of the Internet is, on top of 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 behaviour. 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; KOCHEN, 1978), written in the 1950's[4].

In recent years, however, with the popularization of OSNs, this research subject gained new momentum as novel possibilities of study have arisen and plenty of data on social relations and interactions have become available. Even though the

[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.

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.

2.1 Online Social Networks as Object of Study

2.1.1 What Attracts Scientists to Study Online Social Networks?

The degree of attention that the media and general public give to OSNs can be a good motivation for the research in this field. However, from a computational perspective, OSNs present some particularities that must be taken into account, in order to understand the 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]. This abundant data is unprecedented in the study of social systems and serves as a base for computational analysis and scientific work. Due to its abundance, social data can fit in the context of *Big Data* research.

Multiple authorship: Differently from other corpora, the textual content produced in OSNs has different authorial sources. This enhances the data collected, which present 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 that composes such networks is an agent able to take decisions and interact with other agents. This complex interactions dynamics produces effects that puzzle and interest several researchers.

[5] <https://myspace.com>

[6] Specified privacy limits and download rates may be imposed.

Temporal dynamics: The fact that social data is generated continuously along time allows analysis that take into account processes and transformations, as topic evolution or collective mobilizations.

Instantaneity: Besides the continuous generation, the social data is also provided at every moment, instantaneously. Thus, OSNs typically reacts 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 and tablets), OSNs content can be generated, virtually, from anywhere at anytime. Also, data's *geolocation*, a feature present in many OSNs, adds new possibilities to the analysis.

2.1.2 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 access to its data.

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; 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 a source.

Twitter, a popular microblogging tool (CHEONG; 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, some time after the beginning of the service, rate

[7] <https://dev.twitter.com>

[8] In Twitter, users can post only up to 140 characters text messages, unlike Facebook, where users can send photos, videos and large text messages.

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. Examples: Guo *et al.* (2011), Qu *et al.* (2011), Yang *et al.* (2012) and Bao *et al.* (2013).

Other web services that integrate social networking features have been the focus of study. Examples are media sites like YouTube[10] (MISLOVE *et al.*, 2007) and Flickr[11] (CHA *et al.*, 2009; KUMAR *et al.*, 2010), and news services as Digg[12] (HOGG; LERMAN, 2009; WU; HUBERMAN, 2007). Research was also made with implicit social networks as email users (TYLER *et al.*, 2005), university pages (ADAMIC; ADAR, 2003; ADAMIC; ADAR, 2005) or blogs (GRUHL *et al.*, 2004), even before the creation of social networking services.

2.2 Categories of Study

In order to simplify the presentation of the wide range of works devoted to the analysis of Online Social Networks, from a computational perspective, 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: (1) structural analysis, (2) social data analysis, and (3) social interaction analysis. Fig. 1 illustrates this proposed structure, with its respective subdivisions.

[9] <http://weibo.com>
[10] <https://www.youtube.com>
[11] <https://www.flickr.com>
[12] <http://digg.com>

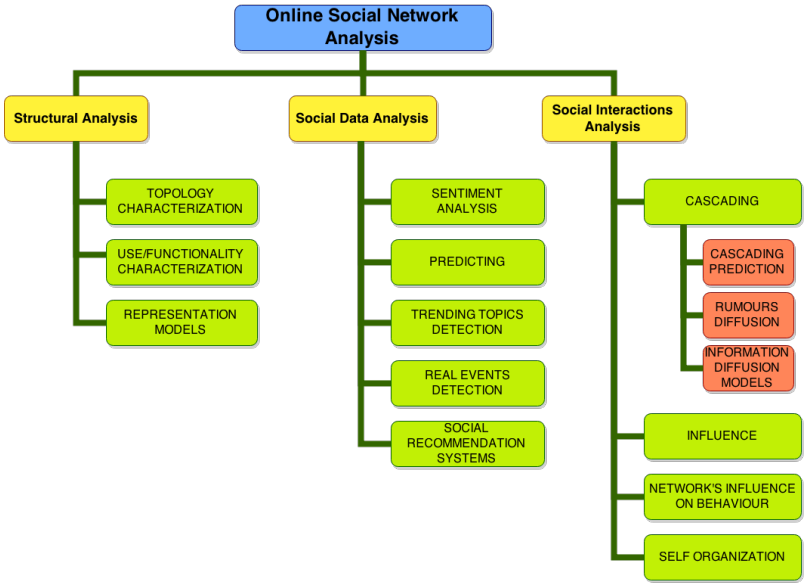


Figure 1 – Categories of study on Online Social Networks, from a computational perspective.

2.2.1 Structural Analysis

Under structural analysis are works that have OSNs structure and operation as objects of study. Many can be the reasons why 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.

From the end of the last century, studies showed that many real networks have some non-trivial properties, such as small average distances between nodes (WATTS; STROGATZ, 1998) and number of connections per node following a power-law (BARABáSI, 1999), culminating in the rise of a new area of study named complex networks or network science (MITCHELL, 2009). Such networks can be found on many areas (COSTA *et al.*, 2007a), 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.

2.2.2 Social Data Analysis

The focus of social data analysis is essentially the *content that is being produced by users*. The data produced is the most important aspect considered, making other features like the social connections secondary.

Although works belonging to this category can be conducted with other datasets (even datasets without social characteristics), the data produced in social networks are rich, diverse and abundant, which makes them a relevant source for data science. 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.

2.2.3 Social Interaction Analysis

Works on social interaction analysis tend to focus primarily on how users are connected (by observing its social connections) and how the data produced by users relate to the network's structure.

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 questionings 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.[13]

[13] See (KURKA *et al.*, 2015) for a comprehensive list of papers on each category.

2.3 Analysis of Spatio-Temporal Phenomena

Many aspects of social behaviour have been interesting for researchers. Studies intended to predict public opinion (ASUR; HUBERMAN, 2010; GRUHL *et al.*, 2005; TUMASJAN *et al.*, 2010), analyse collective sentiment (LANSDALL-WELFARE *et al.*, 2012; STIEGLITZ; DANG-XUAN, 2012) and investigate the formation of structures (BOLLEN *et al.*, 2011; KWAK *et al.*, 2010) are some examples. However, as already mentioned in Section 1.2, the focus of the present work is on the investigation of how information is disseminated and processed throughout social networks.

Those events can be characterized as spatio-temporal phenomena. As OSNs are distributed systems, both spatial aspects (such as the network topology) and temporal (such as speed or synchrony) should be considered in order to observe and study information diffusion processes.[14]

Extensive research has been carried out in order to model the propagation of information in OSNs (BORGE-HOLTHOEFER *et al.*, 2013). Models originally used in other fields such as medicine, biology and physics, can be used to describe this effect (DRAIEF; MASSOULI, 2010). However, given the inherent complexity of the various factors that govern this phenomenon, there is still much research to be developed and implemented, what gives space and relevance to the present work.

Attempts to predict the spread of messages in networks have also been made, however with little success (CHEN *et al.*, 2010; LEE *et al.*, 2014). Prediction normally considers three factors (AARTS *et al.*, 2012), not necessarily simultaneously: the message (content) propagated (e.g. Romero *et al.* (2011)); properties of the users involved in the diffusion (e.g. Borge-Holthoefer *et al.* (2012), Suh *et al.* (2010)); and network structures in which content is diffused (e.g. Ardon *et al.* (2013), Lerman e Ghosh (2010)).

Research has also been done investigating the power of influence of individual

[14] The term "spatio-temporal" might be used in a more restrict sense, referring to networks where the physical location of its users along the time is accessible and relevant (see Shekhar e Oliver (2010) and De Longueville *et al.* (2009), for example. This work, however, opts to use the term in a wider sense, as explained in the paragraph.

users (BORGE-HOLTHOEFER *et al.*, 2012; CHA *et al.*, 2010) and characterizing different patterns in the dissemination of false and true topics (CASTILLO *et al.*, 2011; TRIPATHY *et al.*, 2010).

2.3.1 Self-organization in networks

Another interesting aspect of social networks is the presence of self-organization. Even without a central control, connected individuals are able, in certain circumstances, to organize and act globally. Collective information diffusion processes are examples of this, since many elements of a network work together (sharing information), producing global effects (the dissemination and popularization of information).

However, other self-organizing processes are present and observable. An example is the network mobilization during crisis events such as social revolutions (GONZALEZ-BAILON *et al.*, 2013; STARBIRD; PALEN, 2012) or natural disasters (VIEWEG *et al.*, 2010). An impressive level of coordination is detectable in emergency situations where authority roles are spontaneously assigned to specific users, by other users, and the network is able to select relevant and useful content in the midst of thousands of transacted messages (SARCEVIC *et al.*, 2012).

Another spatio-temporal phenomenon associated with self-organization is related to how the network selects and filters opinions and information. Salathé *et al.* (2013) explores the scattering dynamics and the discussion of favorable or contrary opinions to a new vaccine. The ways in which rumors gain or lose space on the network is also a subject of study by several researchers (CASTILLO *et al.*, 2011; GUPTA *et al.*, 2012; MENDOZA *et al.*, 2010; TRIPATHY *et al.*, 2010).

Despite several of the cited studies have already approached similar themes to those that form the basis of the proposal of this research, the present study differs from the others by proposing an extensive use of the complex network theory to characterize and classify the observed events of diffusion and information processing. Although some studies use concepts of the area, there is still a stronger emphasis on the textual content of the information being spread and not on the structure and the

propagation dynamics intrinsic to the content.

3 Complex Systems

Different approaches can be used, when studying a process, in order to build explanatory models. In this context, reductionism has been one of the most common paradigms, proposing the progressive division of an original problem in simpler instances with known solutions, having the general solution as a conjugation of the problem's partial solutions (DESCARTES, 1637).

Many problems assessed nowadays by science, however, challenge such a paradigm (KUHN, 1962). Issues like weather forecasting, the study of an economic system, or epidemics propagation have in common the fact that, when partitioned in manageable elements, lose relevant information, compromising the validity of a general solution. Non-reductionist approaches are required to explain them, as it becomes clear that, in some cases, "the whole can be more than the sum of the parts" (AMARAL; OTTINO, 2004).

When studying a system from this perspective, the relationship among the components and how they mutually affect each other should also be explored, additionally to the emphasis on its individual features. The understanding is that, from such mutual relations, the system might present a novel (or emergent) dynamics, which is not present in or even achievable by the individuals.

Systems that require such an special generalized approach are called *complex systems*. Complex systems are present in many fields such as physics, mathematics, biology, and social sciences. Although a closed definition for the term may not be generally accepted, some characteristics are commonly present in all such systems (MITCHELL, 2009):

- Presence of multiple individuals with autonomy to take decisions, without the control of a central leader. The absence of a central control does not imply that the individual decisions are going to be independent, due to the existence of an

intricate network of cause-effect relations. The complexity can be understood as a consequence of those properties, so as when the number of individuals increase, the number of individual decisions and the possibility of mutual interference among them also increase, creating a virtually unpredictable, and many times interesting, behaviour of the system.

- Communication between connected individuals. The ability to exchange information allows the occurrence of processing in the system, as information and signals are able to travel throughout the system and influence individual decisions.

- Adaptation by means of evolution or learning. Complex systems are able to adapt, in reaction to environmental pressure, having flexibility to survive even under varying environmental conditions.

Therefore, the complex systems approach is very appropriate for the study of online social networks, when observing phenomena that progress from individual to global behaviour.

One of the sub-topics of the complex systems theory that is more adequate to the social network analysis is the *complex network analysis*, that is the study of network interaction structures that possess specific topological features.

3.1 Complex Networks

The study of complex networks uses graph theory to model and represent structures from the real world. The basic components of a network (or graph) are vertices and edges that connect them. On a network representation, vertices correspond to individual units, like neurons in a neural network or human beings in a social network, and the connections correspond to their relations (synapses or friendships, respectively, in the given examples).

The connections tend to have a crucial role in the network's local and global behaviour, although the vertices may have their own properties that can be studied and analyzed separately, as connections allow the exchange of material, energy or information between its elements. The individual elements of the network, thereby, can surpass local limitations. This is clear, for example, in human being networks: the person's connections can influence him or her to take decisions or acquire knowledge that may go beyond his or her initial individual potentials (BARRETT *et al.*, 2007; HUTCHINS, 2000). When the amount of individuals and connections increase, the complexity of the processes also increase, bringing emergent behaviours that cannot be independently assigned to any of the individuals.

3.2 Complex Networks Characterization

3.2.1 Graph Theory Definitions

A graph is defined by its constituent parts: **vertices** and **edges**. Therefore, a graph $G = \{V, E\}$ can be characterized by a set of vertices V and a set of edges E. While the vertices in V are independent elements, the edges in E are defined by a pair of nodes in V which are connected.

The components from both sets (V and E) might have attributes, depending on the application. For example, a vertex might have labels and other attributes, representing properties such as weight, power or name. An edge might have a real value associated with it, representing the intensity (or cost) of a connection.

Edges can also be directed, representing a precedence between two vertices, thus characterizing a *directed graph*. When the edges represent merely a connection between two vertices, without direction, the graph is called *undirected*. Directed edges are graphically represented by arrows, and undirected edges by lines (see Figure 2.

If set E contains more than one pair $(i, j) \in V$, the repeated edges are called *multiple*. When it is possible to find a sequence of edges in E that connects two vertices i and j, possibly passing through other vertices, a *path* between i and j is

defined. A path that starts and ends in a same vertex is called a *cycle*.

Some specific cases of graphs receive its own nomenclature:

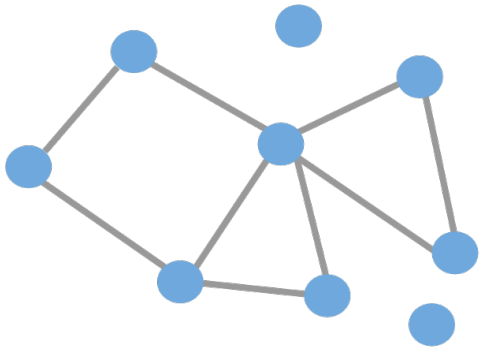
- **Simple graph:** Graphs without cycles or multiple edges;
- **Directed acyclic graph (DAG):** Directed graph without cycles;
- **Complete graph:** A graph where the edges in E connect all possible pair of nodes in V;
- **Connected graph:** A graph where for every pair of nodes (i, j) there is a path in E.

A graph can be represented in several ways. Graphical representations usually describe vertices as circles or dots and edges as lines (or arrows if directed) connecting them. Mathematical representations include the definition of set V as a list of elements and E as a list of vertex pairs. Another useful notation, specially used for arithmetic operations, is the matricial notation. Examples of the use of matrices to represent graphs include the *incidence matrix*, where rows represent the graph's vertices and the columns the edges and the cells define if a specific vertex (row) is present in a specific edge (column). On the *adjacency matrix* both rows and columns are indexed by the vertices and each cell represent the existence of an edge connecting the pair of vertices defined by its (row, column) position.

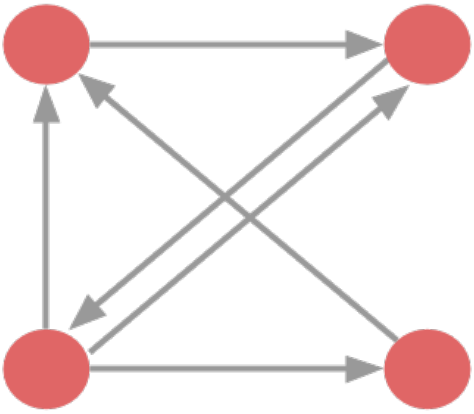
3.2.2 Network Metrics and Properties

Several metrics can be used to characterize a given network, represented mathematically by a graph with edges and vertices, $G = \{V, E\}$. The aim of such a characterization is to analyse the network's properties and also to be able to compare different network structures.

There are many metrics used by researchers, as shown by Costa *et al.* (2007b). The ones most relevant to the present work are detailed below:



(a) A simple undirected graph.



(b) A directed, connected graph.

Figure 2 – Examples of graphs.

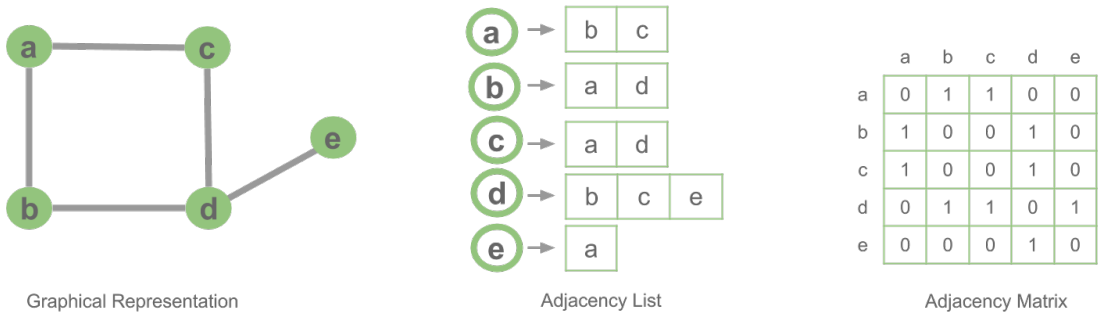


Figure 3 – Different representations of the same graph.

Size and order: The total number of edges and vertices of a graph G are noted respectively $|E|$ and $|V|$. $|V|$ is also called the graph's *order* and $|E|$ its *size*.

Network's density: The ratio between the total number of edges $|E|$ of a network and the maximum number of edges defines the network's density D. For an undirected network, we have:

$$D = \frac{2|E|}{|V|(|V|-1)}$$

Average degree: The total number of connections of a vertex i is called vertex's *degree* and is denoted by k_i. In a directed network, there are two types of degree measurements: *in-degree* (k_i^{in}) and *out-degree* (k_i^{out}), defined as the amount of edges incoming or exiting a specific vertex. A general degree measurement can be defined, in directed networks, as $k_i = k_i^{in} + k_i^{out}$.

This index is relevant, as the vertices with large number of connections are denominated *hubs* and are major intermediaries in the flow of activity or information of a network (MITCHELL, 2009).

It is possible to determine a network's average degree $< k >$, from the calculation of individual degrees for each vertex of a network, as:[1]

[1] Note that, in directed networks, $< k^{in} >=< k^{out} >$.

$$\langle k \rangle = \frac{1}{|V|} \sum_i k_i = \frac{2|E|}{|V|}$$

Degree distribution: Another useful metric, using nodes' degree is the *degree distribution*, which is a histogram, counting the number of vertices in the network for each degree k. When the histogram is normalized by the number of vertices $|V|$, a probability density function (PDF) can be derived, giving the probability $P(k)$ of a random node have degree k.

Distance metrics: A path's *length* is defined by the number of edges in a path connecting two vertices. The length of the shortest path (or paths) connecting two vertices i and j is called *geodesic distance* ($d_{i,j}$). If there are not paths connecting i and j, the geodesic distance is defined as $d_{i,j} = \infty$.

After computing the geodesic distance between all possible pairs of nodes in a network, two relevant metrics are found:

- Average distance[2]:
$$L = \frac{1}{|V|(|V|-1)} \sum_{i \neq j} d_{ij}$$
- Network diameter: The longest geodesic distance found in the network.

Distance metrics are useful for measuring, in average, the number of steps it takes to move from one member of the network to another.

Clustering coefficient: a network's cluster is defined by highly connected groups of vertices. To quantify how clustered is a vertex or a network, a clustering coefficient is defined. An evidence of a cluster is the occurrence of "triangles"

[2] When the network is not connected, some geodesic distances are equal to ∞ and the average distance diverge. Alternatives to avoid this case is to consider only the distance of pairs belonging to the same connected component, or the use of a related measurement, as the *harmonic mean* proposed by Marchiori e Latora (2000), defined as:

$$L' = \frac{1}{L} = \frac{1}{|V|(|V|-1)} \sum_{i \neq j} \frac{1}{d_{ij}}$$

– a cycle of three vertices, fully connected. Therefore, for a single vertex i, its clustering coefficient is calculated as:

$$C_i = \frac{N_\Delta(i)}{N_3(i)}$$

Where $N_\Delta(i)$ is the number of triangles involving vertex i and $N_3(i)$ is the number of pairs of nodes connected to i.

If l_i is the number of edges between any pair of neighbours of node i, and k_i, the degree of node i (i.e. its number of neighbours), C_i can be expressed on an undirected network as:

$$C_i = \frac{2l_i}{k_i(k_i - 1)}$$

For the whole network's clustering coefficient, two metrics are possible:

- The mean of all individual coefficients:

$$C = \frac{1}{|V|} \sum_i C_i$$

- The total number of triangles in the network (N_Δ) divided by the number of present connected triples (N_3):

$$C = \frac{3N_\Delta}{N_3}$$

Node centrality: The position of a node in a network can interfere on its access to information and its participation on events that take place on the network.

Node centrality metrics measure the importance of a vertex in the network. There are several metrics currently used. One of them is the already mentioned node's degree.

Other possible metrics is the *betweenness centrality*, that measures a node's centrality by how often it appears in paths connecting occasional pairs of vertices. Quantitatively it can be defined as:

$$b_i = \sum_{j,k} \frac{\sigma(j, i, k)}{\sigma(j, k)}$$

where $\sigma(j, i, k)$ is the number of shortest paths between vertices j and k that passes through vertex i and $\sigma(j, k)$ is the number of shortest paths between j and k, without restrictions.

Finally, another example of metric is the *eigenvector centrality*, that in order to measure a node's centrality, consider not only its degree, but also the degree of its neighbors. Therefore, a node connected to high degree nodes is considered more central than another node connected to the same amount of neighbors, but with smaller degrees. An adaptation of this metric was used and popularized by the search engine Google, in the *PageRank* algorithm (PAGE *et al.*, 1998).

3.3 Special Properties of Complex Networks

3.3.1 History of Developments

The consolidation of complex networks as a field of study is a recent issue. Until a few years ago, due to the lack of technological resources and theoretical tools, network relations that were present in the real world were considered as the result of chance.

An early network formation model that captures this concept is the one proposed by Erdös and Rényi (ERDöS; RéNYI, 1959), where networks are modeled from random graphs. In their model, total number of vertices in the network is first defined. Then, edges are randomly placed connecting the initial vertices. Therefore, any network with a given number of individuals and connections have the same chance of being generated.

The development of more powerful computers and the greater availability of data for analysis increased the knowledge on real networks. However, when those empirical networks were analysed, it was shown that Erdös–Rényi model was not a proper solution to reproduce the behaviour of the observed networks.

Real networks presented nontrivial properties that were not part of the purely random or regular models. By the end of the twenty century, studies that indicated

some of those peculiarities gave great impulse to the research on complex networks. Three relevant discoveries can be highlighted and will be detailed in the following sections: *small-world* property (WATTS; STROGATZ, 1998), *scale-free* models (BARABáSI, 1999) and the identification of community structures (GIRVAN; NEWMAN, 2002).

3.3.2 Power Law Degree Distribution

A simple way to characterize a network is through the analysis of its *degree distribution*, $P(k)$. In a random network, like in Erdös-Rényi model (ERDöS; RéNYI, 1959), it is expected a normal distribution of degrees, where the majority of vertices have degree near the average $< k >$ and as the degree deviates from the mean (above or below it), less vertices are found.

The observation of several real networks, however, showed that typical network's degree distributions generally followed a power-law, of form $P(k) \propto k^{-\gamma}$, where γ is a positive constant (BARABáSI, 1999). In this distribution, most of the network's vertices have small degrees, while few vertices have high degrees of connectivity. Also, on fat tail distribution (that is the case for power-law) more and bigger hubs are present than on Gaussian networks.

As the power-law distributions are invariant to scale changes, networks with this property are also commonly called *scale-free networks*. The scale-free networks present some remarkable properties (MITCHELL, 2009):

- The presence of a relatively low number of vertices acting as hubs, concentrating a large number of connections. The analysis of the information flow on networks show that those hubs are more likely to have contact with relevant information and also greater capacity to disseminate information;

- Heterogeneity on degree values: the vertices of the network have degrees that cover a wide range of values, presenting diversity on individual vertex's roles and "influence";

- Self-similarity: parts of the network have similar properties to the complete network. This feature is also characteristic of fractal structures;

- Resilience: even with the removal of random vertices, there is a tendency to the network's original connectivity properties be preserved (unless in the unlikely case where the vertices removed are high degree nodes). This implies the network ability to maintain its operation, even with a significant removal of vertices.

Previous studies (BARABáSI, 1999) show that social networks generally have power-law degree distribution. Intuitively, the idea is plausible when analysing society: while few have a great number of relationships (corresponding to a high connection degree), most part of people have a relatively low number of relations. And, in a new social environment, the chance of an individual knowing another already "popular" person, with many connections, is higher than the chance of knowing people with less connections, isolated (maintaining then the scale-free structure of the network).

3.3.3 Small World Effect

One of the most noted properties observed on real networks is the small world effect.

The name refers to Stanley Milgram's work (MILGRAM, 1967), who analysed how people were connected in society and the possibility of finding short paths between individuals. From the beginning of the twentieth century, there was a popular conjecture stating that it was possible to connect any two persons by means of an average of five intermediates, also known as "six degrees of separation" (KARINTHY, 1929).

In order to empirically validate this notion of proximity in society, Stanley Milgram proposed an experiment where a group of people living in midwest United States received instructions to deliver a letter to a target person living in Boston,

Massachusetts (considered far both geographically and economically). The instructions given to the first recipients of the letter were to either: (a) send the letter directly to the target, if they knew the target personally, or (b) chose a known person that, in their opinion, were more likely to know the target.

The paths took by the letters were registered and analyzed. From the letters that reached their recipient (there were many cases where people refused to keep the experiment, breaking the chain), it was possible to calculate a number between five and six as the average path length, showing evidences of the "small-world" effect.

Years later, in 1998, Duncan J. Watts and Steven Strogatz (WATTS; STROGATZ, 1998) could generalize Milgram's result, by analysing other real networks and finding two non-trivial properties: small network diameter and high clustering. From the observation, a generative model was also proposed.

It was verified that real networks have a smaller diameter than purely random networks (BOLLOBáS; RIORDAN, 2004). This implies that vertices on a complex network can find relatively small paths between each other. This result allows that disturbances on the network be propagated very fast. This can help to explain the speed and the spatial patterns with which rumors and epidemics spread in society.

The presence of highly connected groups is common in complex networks, such as social networks, forming *clusters*. The clustering coefficient in random networks is inversely related to the number of vertices on the network, so that the increase of vertices tends to separate groups and increase the network's sparsity. On small-world networks, however, it was observed that the clustering coefficient C remains constant, independently of the number of vertices. Large and complex networks thus tend to have a much higher clustering coefficient than random networks.

Another common feature of social networks, besides clustering, is the presence of communities. Communities are characterized by groups of vertices highly connected among them, but poorly connected externally to other groups. This effect is also intuitively true for social networks, as it is common the identification of concise groups (joined by affinity, age, social aspects, regional proximity, etc), with the

described properties.

In terms of information flow, groups with highly connected vertices may imply a redundancy of communication channels, possibly enhancing information that pass through or are generated by these groups. In the case of online social networks, it can imply that content introduced on those clusters is proliferated and reinforced. It is expected also a community influence on how its members receive and judge information.

3.3.4 Relation between structure and function

Another aspect in what complex networks differ from general or random networks is in the presence of adaptive structures.

It is possible to observe recurrent patterns of connections among vertices in real networks, called *motifs* (SHEN-ORR *et al.*, 2002). From the theoretical analysis of random networks, it is possible to statistically predict the chance that pattern of connected vertices (sub-graphs) appears in networks with characteristics – like degree distribution – similar to those of real networks. However, it is empirically observed on real networks that some sub-graphs appear in a much higher frequency than expected, defining thus a motif. Figure 4 presents some common motifs examples, found on known networks.

This high frequency may indicate that specific functions of the network are being assigned to these groups of vertices (MILO *et al.*, 2002). Evidence of this phenomena can be seen, for example, in information processing networks like neural networks and gene transcription networks, where feed-forward loops (see Figure 4) act as filters, generating pulse and response accelerators (ZASLAVER *et al.*, 2004).

A relationship between the topological structure of a network and the behaviour of its components can also be often noticed. In most cases it is not possible to determine which is the cause and which is the consequence (i.e., if the topology is a result of individual behaviour, or if the behaviour is a consequence of the topology), but the study of one can generally help understanding the other.

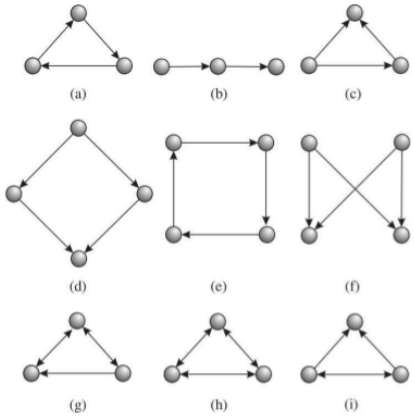


Figure 4 – Some typical motifs and their denominations: (a) three-vertex feedback loop, (b) three chain, (c) feed- forward loop, (d) bi-parallel, (e) four-vertex feedback loop, (f) bi-fan, (g) feedback with two mutual dyads, (h) fully connected triad and (i) uplinked mutual dyad. Image from Costa *et al.* (2007b)

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 Online Social Networks. For example, Bollen *et al.* (2011) verified, by investigating the relationship between emotions and social connections, that users considered happy tend to be linked to each other.

3.4 Social Networks as Complex Systems

Now we return to our main object of study – Online Social Networks – and their relation to complex systems. The aim is to determine in which sense social networks can be considered complex and how can social systems be modeled in this framework.

3.4.1 Modeling Social Networks

An immediate approach, when modeling social systems as networks, is to consider each individual as a vertex and its social connections as edges. This representation suitably models most online social networking services available, where each person has a user profile and is able to connect to other people, as their friends or followers.

Attributes can be used to specify characteristics from both users and connections. For example, each vertex of a social network model may have associated with it a profile with personal information about the user it represents and groups of interest he or she shares with other users of the network. Edges can also have attributes, modeling aspects such as "relationship intensity" or geographical distance.

From the network's structure, it is possible to infer some relevant aspects about the individuals, such as presence of communities, preferred relationships and paths of communication, and information's flow. Issues like influence, relationship building and collective movements can also be noticed, if analysed over time.

3.4.2 Multiples Networks in a Network

Different ways of modeling the system are also possible and can bring different perspectives to the analysis. Some alternatives are highlighted:

- Instead of modeling the network's edges using the "declared" list of relationships on users profiles, it is possible to infer connections from user's interactions. This implicit network can reveal interesting aspects about user's relationships and serves as an alternative for a more reliable social analysis.
- Characterizing vertices as group of users, instead of individual users. This would favour the analysis of groups interaction and collective dynamics.
- It is common that a group of individuals belongs to more than one network, in different contexts. For example, the same group of people might have accounts in a friendship social network (e.g. Facebook), a professional network

(e.g. Linkedin) and a news social network (e.g. Digg). The overlap of all those network structures can create a *multiplex network* (GOMEZ-GARDENES *et al.*, 2012; GÓMEZ *et al.*, 2013; MUCHA *et al.*, 2010; SUN; HAN, 2012), allowing new discoveries.

- Inclusion into the model of different objects as vertices, alongside the users (SUN; HAN, 2012; BARBOSA *et al.*, 2015). For example, in a movies social network, vertices could represent either users or movies. Thus, connections between movies and users would represent movies previously watched and connections between users would indicate friendships.

3.4.3 Complex Behaviour

Complex systems and complex network theory provide a general framework for working with many different systems, in different areas, as seen in the sections above. The same graph structure can be used, for example, to represent the interaction of proteins in an organism, atmospheric agents in weather forecasting, pages and hyperlinks on the Internet or even a country's electrical distribution system.

The same theory applied to human relationships, however, seems to be controversial when reducing a whole person, with all its uniqueness, to a simple and homogeneous vertex in the network. It is interesting to notice that one of the premises of complex systems is that the complexity does not come from the individual agents alone, but from the collective experience and the multiplicity of actions and their intertwined effect. Thus, although each individual of the system can be seen as a whole "complex system", this is not the complexity sought in this work.

This tension between the power of the individual behaviour versus the influence of an emergent force, created by the collectivity, is a topic of author's interest, and this study intends to bring questions and initial answers to this topic.

What is the capacity of an individual to resist the influences of his or her environment and keep an opinion or a behaviour? Is it possible to characterize indi-

viduals by characterizing a system? To whom does the values and beliefs propagated over the network correspond?

Those are questions that certainly will not be totally answered in the scope of the present work, but they will guide the investigation behind the experiments of the next chapters.

4 Experiments and Methodology

In the following chapters and sections, experiments conducted in order to validate and propose new perspectives to Online Social Network analysis are presented. Real data, extracted from social networking services, were used and a methodology of experimentation was developed to validate the hypothesis.

This chapter present details of the proposed methodology and the key decisions taken to accomplish practical results.

4.1 Questions to Be Answered

There is a general question that permeates the experiments described in the following chapters:

Main question. *Is it possible to explain individuals behaviour, from external observations of the topology and collective behaviour?*

In order to seek possible answers to this question, two lines of experimentation were conducted, each addressing a different aspect:

- Experiments set I: Investigation of information diffusion's properties, in order to find patterns and predict features of the content being transmited over the network (Chapter 6);
- Experiments set II: A study of the interplay that social connections may have on users' behaviour over the network (Chapter 7).

Each set of experiments has also a fundamental question and a practical question that guide the process of analysis. The fundamental question considers philosophical implications behind the proposed experiments. The practical question

presents a formulation of the fundamental question with terms that can be concretely answered by computational experiments and techniques.

Thus, from the direct computational experiments, it is expected that hints of the fundamental and practical questions can be revealed and explored.

4.2 Choosing a Social Networking Service

In order to analyse real data, a social networking service had to be chosen. A service that provided significant amount of data, diverse and generic enough to allow the formulation of general theories and suitable to different contexts and observations was aimed. After evaluating possibilities, the choice was made for the *microblogging* service *Twitter*[1].

In this service, individual users can own accounts and are able to associate their accounts to others users accounts (*following*, or *being followed* by other users). Each user can also publish texts up to 140 characters named *tweets*, accessible and visible to all their "followers", see Figure 5. The service also provides an interface where each user can see a *timeline* that contains a list of the content posted by all the users they follow, in chronological order.

A *tweet* can be an original message authored by a user, or it can be a replication of another message previously created by another user. The latter case is a popular practice on the network and is called *retweet*. A retweet can be identified by a message prefix composed of the letters "RT". Currently, the process of *retwitting* can be done manually (by copying the original message's text and publishing a new message, prefixed by "RT") or via the official service's interface, that provides a quick button to reshare a specific message with the user's *followers*.

Another popular convention, created initially by the users, but later officialy adopted by the service, is the use of *hashtag* on tweets, with the purpose of labeling a message ("folksonomy"). A hashtag consists in a string of characters (generally a

[1] http://twitter.com/

Figure 5 – An example of tweet.

word or a small phrase without spaces), prefixed by the hash ('#') character.

The presence of retweets and hashtags is very helpful, as it allows the easy identification of information cascades throughout the network, by analysing users retweeting the same message, and the formation of topics, by analysing the hashtags used to describe a group of messages.

Also, *Twitter* presents other beneficial features that justify its choice as a database:

- There is a clear representation of the individuals (users) and their relationships (followers and followees), allowing the characterization of a network;
- There is a regular way of interactions between the users: short messages with 140 characters;
- The information that travels throughout the network is easily treatable, since

it is restricted to textual content[2];

- There is a great amount of public data available, produced by users;
- The existence of APIs enables the access and manipulation of data;
- Literature in the extraction of data in virtual networks using *Twitter* is consolidated.

The App Programing Interface (API) provided by *Twitter* [3] was used to obtain data. Two different interfaces were used: (a) the *Rest API*, to make requests on Twitter's database, extracting data regarding users profiles and their relationships in the network, and (b) the *Stream API*, to filter in real time public messages of interest. Both APIs have limits for the allowed requests. It was possible, however, to extract enough data for the research conducted here obeying the limits.

4.3 Methodological Framework

4.3.1 Data Acquisition and Modeling

The APIs described above were used to extract data from the Twitter service to build models of information diffusion, that could be further analysed and interpreted.

One restriction imposed by the API's rate limits is that the number of requests to get messages posted on the past is significantly low, making significantly difficult the analysis of already documented diffusions. The alternative found was to use the *Stream API*, that allows the extraction of a vast quantity[4] of messages that are being transmitted in real-time on the service.

[2] Although recent changes on the official service interface (<https://twitter.com/>) interpret the content of urls on messages, highlighting the presence of references to photos or videos.

[3] http://dev.twitter.com

[4] The Stream API allows the extraction of up to 1% of all the messages published on the service, during its use. Considering that Twitter is a global service, this limit is, in practice, quite permissive, allowing the extraction of full cuts of the network's activity.

Although the amount of data extracted was not an issue when using the Stream API, the necessity of using real-time data imposes restrictions on the diffusions that can be analyzed. The main challenge was to determine *a priori* the diffusion processes that are going to happen in the near future and how and where a diffusion process will begin, in order to collect its information.

The solution found was to focus the analysis on diffusions processes created and started by popular users on the network. Popular Twitter users, having many followers, are able to start cascade of messages (through retweets), from the content posted by them (tweets). Therefore, by filtering messages authored by famous users (such as actors, personalities, public figures or news agencies) and their shares, it was possible to witness in real-time several processes of information diffusion with varying sizes.

From the observation of the occurrence of such processes, it was possible to map the users that took part in the diffusion processes, their relationship network and the time instant in which the messages were sent. As a result, complex networks could be synthesized, describing the information flow observed.

4.3.2 Analysis of Features and Interpretation

The modeled networks could then be analysed. In this stage, metrics and parameters were searched, to extract a set of properties and spatio-temporal attributes for the database obtained. The properties observed involved general network characteristics, such as average connections, size, distances and spatio-temporal dynamics. Details of complex networks characterization can be found in Chapter 3.

Beyond the characterization of the metrics obtained from the network, it was possible to do a qualitative analysis of the diffusion processes, analysing aspects such as the subject of the messages that were being spread.

Finally, using machine learning techniques (BISHOP; NASRABADI, 2006; MITCHELL, 1997) involving classification, regression and data clustering, it was possible to discover patterns and relations in the database.

5 Database Description

This chapter will describe the database extracted from *Twitter's* service and used in the experiments, according to the methodology presented in Chapter 4. It will be shown in the following sections that the data collected presents a rich collection of social activity, enabling the study of complex networks with thousands of vertices and edges.

5.1 Data Extraction and Selection

A source of news to be monitored, *Folha de São Paulo's* account[1] – a popular Brazilian newspaper, with over three million followers, was chosen, with the purpose of obtaining a source of diverse content, that approached different categories, having a wide and varied group of followers.

From March 19 to September 21 of 2014, both the messages (*tweets*) posted by the source, as well as the shares made by other users (*retweets*) were collected. In addition to the messages' metadata (such as time of posting, locality, language, links, *hashtags*), information of the users that retweeted the messages were collected. It was collected personal information (as name, country, account's activity), as well as the list of all their social connections on the network (followers and followees).

It was possible to track, from the data collected, a large amount of information diffusion processes triggered by the observed source. For each process, the characterization of the users' network behind it and the spatio-temporal dynamics of the information flow were registered.

During the observed period, 13463 distinct and original messages posted by the source account were collected.[2] From this set, a series of filters were applied,

[1] <https://twitter.com/folha>

[2] Due to electrical power shortages during the period, the collection process was not uninterrupted.

forming a more appropriate selection to the work, as described below.

- Filter 1 - Original set messages which had received more than 50 *retweets* were selected, resulting in a group of 1487 distinct messages;
- Filter 2 - Messages from six main categories ("everyday news", "sports", "world", "politics", "entertainment" and "market") from the above set were maintained, resulting in a group of 1021 messages;
- Filter 3 - Automated scripts (*bots*)[3] used to automatically retweet every (or almost every) publication were removed from the database and from retweets count (which happened in just one case).
- Filter 4 - As 2014 was characterized by a presidential election in Brazil, the topic "poder" (power), where political news are posted, had a much higher frequency of messages than other categories. Therefore, random messages from that category were removed, leaving a total of 200 messages (from its original count of 329), resulting in a database with a more equalized category distribution.

The filtered data resulted in a collection of 890 messages (M), of which 44571 distinct users (U) *retweeted* one or more messages at some moment, with an average of 95 *retweets* per message. Those data were organized in a matrix of tweets (T) of dimensions $M \times U$, with the element $T_{i,j}$ made equal to 1 in case the message i was *retweeted* by user j, and 0 otherwise (see Figure 6).

From the users lists of followers and followees, it was possible to characterize a graph (G), registering the social connections between them. Among the 44571 users that participated in the information diffusion, 645409 connecting edges were found. The proprietor user, Folha de São Paulo, was removed from the network, so that the interactions among the main page's followers were the focus of analysis.

[3] Users that retweeted more than 80% of the messages were considered as bots.

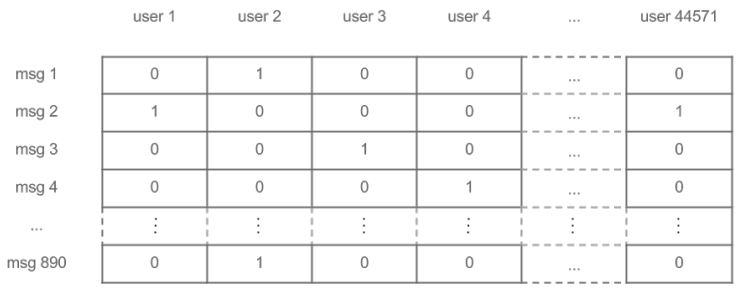


Figure 6 – Graphical scheme of matrix T.

It should be noted that the decision of using only one source of content and the filters applied to the messages and users limited the amount of data used in the experiments. Although the volume of data available on online social services can be treated as *big data* (KAISLER *et al.*, 2013), the database used in this work was intentionally restricted, enabling the application of different computational methods on it.

5.1.1 Classifying the Tweets

Practically all the messages published by the source are headlines of news, followed by a link to the newspaper's website[4], with the news' full content. As the news articles on the website belong to thematic categories (newspaper's sections), it was possible to automatically attribute a class to each *tweet*, based on its thematic category.

This procedure was carried out to all tweets and six predominant topics were verified, among them: "everyday news", "sports", "world", "politics", "entertainment" and "market" (in Portuguese: "cotidiano", "esporte", "mundo", "política", "entretenimento" and "mercado", respectively). Those categories were used on the filtering process (as described above) and were considered as classes on supervised machine learning algorithms, as described in Chapter 6. Table 1 contains the user's distribu-

[4] <http://www.folha.uol.com.br/>

tion according to the number of categories shared.

Although this procedure configures a practical and reliable way to classify the *tweets*, it is possible to question the relevance of the classes. The definition of each topic's class is made by the newspaper's staff, with a specific purpose of organization and separation. Also, news are many times categorized in a subjective way or should be in more than one class, which is not supported by the newspaper.

Alternative classification schemes could have been explored and compared, depending of the goal intended. Such an issue is further discussed in Chapter 6.

5.2 Database Features

The following properties could be verified from matrix T:

Tweets distribution per user: the average user participation, in different published messages, is low. During the observation period, each user retweeted in average only 2.01 messages from the source. Figures 7a and 7b show an histogram of users' posts distribution. The linearity exhibited by Figure 7b indicates a distribution of power law in the user's behaviour.

Tweets distribution per message: each message contains an average of 95.03 retweets, with the most popular message receiving 828 retweets. A distribution profile is noticeable in Figures 8a and 8b. Messages that contained less than 50 retweets were not considered.

T **density:** The total number of messages retweeted is 97033. In spite of matrix T having dimensions 890×44571, its density is of only 0.001975 (i.e. 0.19%), indicating that it is extremely sparse.

Topics distribution: Figure 9 shows the distribution of topics categories (using Folha de São Paulo's classification) on the database. Two important events happened in Brazil and affected the topic distribution: presidential election

Table 1 – User's distribution according to the number of categories shared. Describes the percentage of users that published messages from one or more categories during the period of observation. The analysis considered only the users with 6 or more messages published during the observation period.

Number of categories shared	User's percentage
1	1.3%
2	7.1%
3	22.4%
4	32.1%
5	22.8%
6	14.3%

(affecting the number of political tweets) and the FIFA World Cup (affecting both sports and entertainment messages counts);

Topics per user: it was verified that users that shared posts frequently (6 or more times) tend to share messages of diverse categories. The majority of those users shared messages from 4 different categories, as it can be seen in Table 1. As most users publish content from different categories, there is a tendency of diversification on the content shared per user.

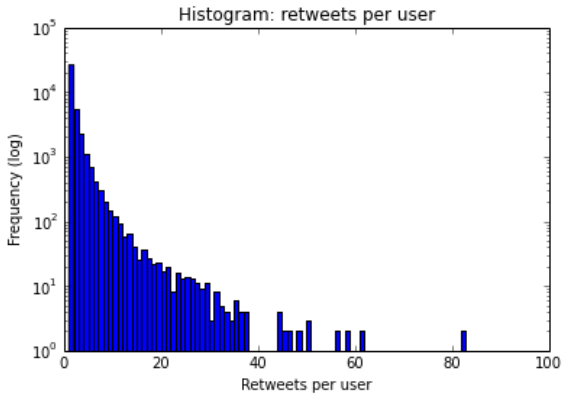
5.3 Network Features

The following properties can be observed from network G:

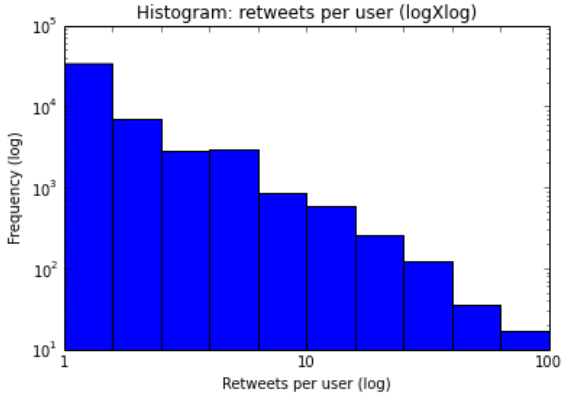
Size and Order: G has 41788 nodes (representing users) and 645409 edges;

Average degree: The average degree $< k >$ of the network is 30.88 connections ($< k^{in} >=< k^{out} >= 15.44$). The user followed by the highest number of people (maximum k^{out}) was the one with the profile of the Brazilian president *@dilmabr*[5], with 5538 followers. The user who followed most other users over

[5] <https://twitter.com/dilmabr>

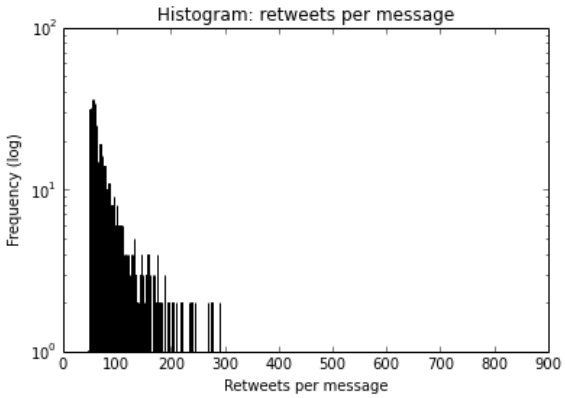


(a) Number of messages shared per user, during the observed period, in a linear × log scale.

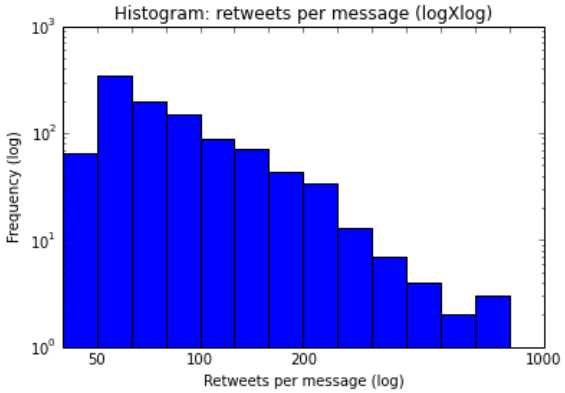


(b) Number of messages shared per user, in a log × log scale.

Figure 7 – Database's histogram of the number of messages shared per user (number of users in log scale). The top histogram presents the evolution of the number of shares (abscissa) in linear scale, while the bottom is presented in logarithmic scale and grouped in larger bins.



(a) Number of shares per message, during the observation period, in a linear × log scale.



(b) Number of shares per message in a log × log scale.

Figure 8 – Histogram describing the number of sharings per message frequency. The figure on top has a linear × log scale and the one at the bottom has a log × log scale. Note that messages with less than 50 *retweets* were removed from the database.

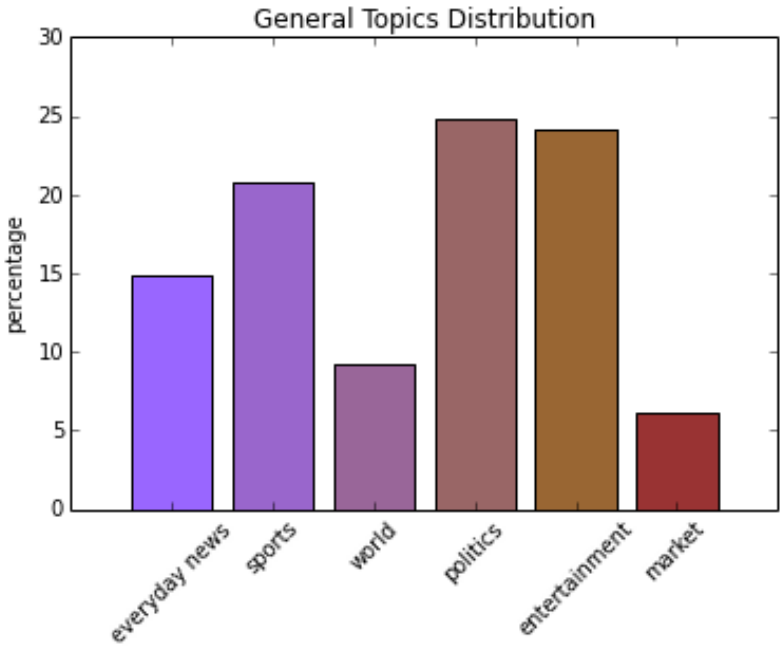


Figure 9 – Distribution of topics for the database.

the network (maximum k^{in}) was *@gilbertotv*[6] account, who followed 2457 users among the ones considered in the study.

Degree distribution: Figure 10 shows a distribution in graph's degrees, indicating the existence of the power law on the out degree's distribution. On the in degree's distribution, it is not so evident, despite the fact that it is also a possible interpretation. The statistical analysis of the curves was made, using *powerlaw*'s python library (ALSTOTT *et al.*, 2014), revealing that both distributions fit on an power-law curve, with $\alpha_{in} = 1.40$ and $\alpha_{out} = 1.53$ as the α coefficients for the in and out degree distributions.

Distance metrics: The network's diameter is 14 and the average path is 4.28.

Components: There are 10968 distinct connected components in total, forming the network, with one of them being a giant component with 30373 users (72% of

[6] <https://twitter.com/gilbertotv>

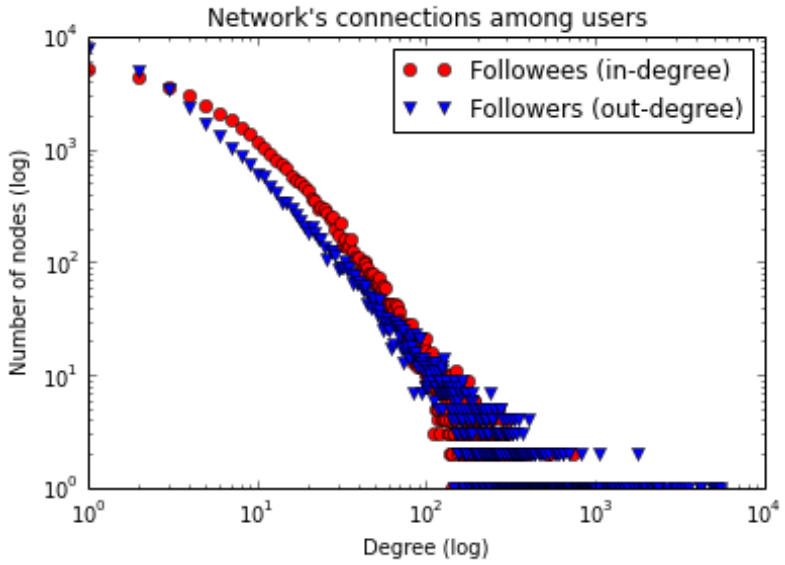


Figure 10 – In and out degree's distribution in the network of connections between users.

the graph's nodes).

Node centrality: The vertex with the largest betweenness centrality was also the one with the largest in-degree: *@gilbertotv*.

Network's clustering coefficient: the clustering coefficient for the whole network is equal to 0.0425, while a random network with the same number of nodes and edges would have an average clustering coefficient substantially lower 0.0005.

5.4 Database's Remarks

Some observations can be extracted from the database based on the quantitative analysis:

Observation 1. *Most users do not participate actively on the diffusions, implying in high diversity of users participating in the processes, but low recurrence;*

Observation 2. *There is no clear consistency in the categories of the messages published by each user;*

Observation 3. *The network observed has properties that are typical of complex networks.*

6 Experiments Set I - Information Diffusion Properties and Patterns

In this chapter, a series of experiments are reported, investigating the structure of information diffusion processes. The aim is to find patterns and predictable features that can be used to understand aspects of the processes taking place in the network.

A fundamental question can be formulated as background to the experiments that will be presented:

Fundamental question 1. *Is it possible to understand the subjective nature of processes happening on a network, from its users behaviour?*

The proposed strategy to answer this question is to use the data collected on each sequence of *tweet* and *retweets* observed to characterize possible relationships between the subject of a message and it's propagation particularities.

Therefore, the practical question that is being addressed is:

Practical question 1. *Given the users that share a message, is it possible to obtain information about the content of the message?*

The content of messages are public and can be obtained directly by its text. However, the experiments presented aim to perceive possible relations between a content published and its effects on the spatio-temporal behaviour of the network. This exploration can give insights on how different content affects people and if this effect can be identified and generalized.

Information of all user's behaviour towards a content can be extracted from matrix T, where each of the M messages is represented by an array M_i with dimension

$U = 44571$. Each array M_i can be interpreted as a "signature" of its associated message.

Machine learning techniques of classification and clustering are used to identify possible patterns and allows a numerical analysis. The techniques used can be grouped in two categories: supervised learning an unsupervised learning.

6.1 Supervised Learning

As the messages analysed can be labeled, using the classification from the newspaper Folha de São Paulo, supervised machine learning algorithms are suitable to be used with the data.

6.1.1 KNN

The first strategy applied was the well-known classification algorithm "*k-nearest neighbours*" (KNN), where a message's category (or class) is defined as the category more present within the k-nearest messages (neighbors) of it.

The distance between two messages is calculated from the behaviour vectors M_i. Diverse metrics were considered to fit to the problem, where the *Jaccard* distance[1] was the one that produced the best results.

Different values of k (neighbours) were tested and the performance is registered in Figure 11. The results presented are the classifier's accuracy for a sample of 300 random tweets of the database. The obtained accuracy is an average of 30 different executions, using different samples.

As a matter of comparison, results from a random classifier were also generated, creating a null model. Instead of using the class of the k-nearest message vectors of a given vector M_i, the random classifier considers the category more frequent within k random messages on matrix T, to determine the class of message i.

[1] Metric used to compare Boolean vectors

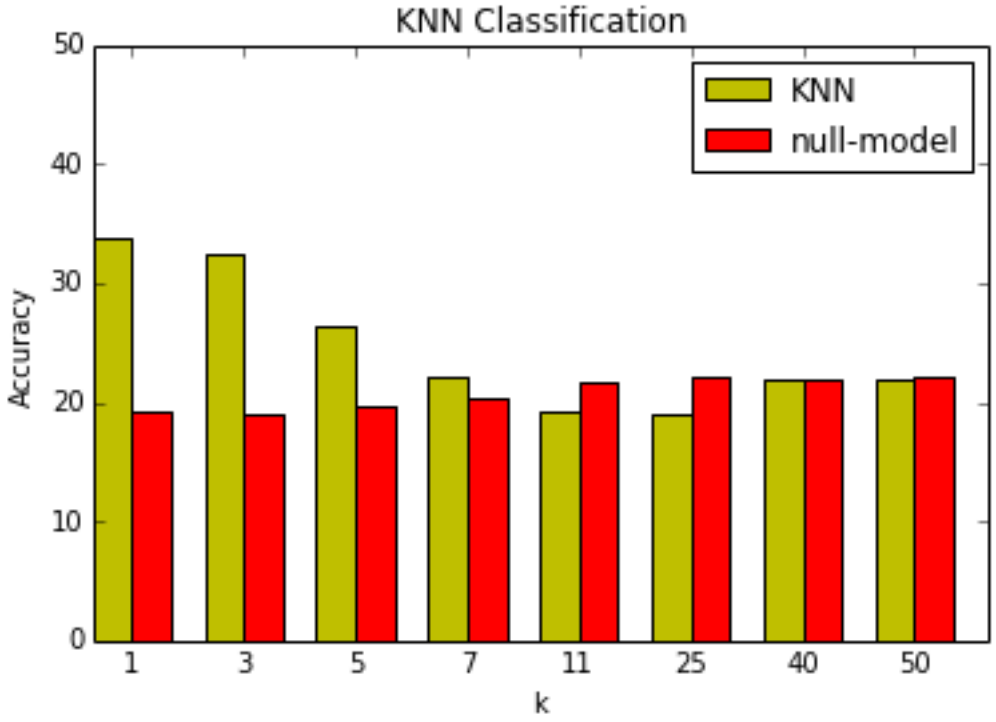


Figure 11 – Clasification results of the *KNN* algorithm and of a random classifier (null model)

It is noticeable that, for lower values of k ($k \in 1, 3$), the KNN classifier shows results way above random, with over 30% of accuracy, showing a relationship between type of message (topic) and participants. For higher values of k, however, the results tend to be equal to the null model.

A possible explanation to this is that each class is formed by small subclasses, spread along the space. Therefore, the first few closest neighbors tend to belong to the same class, but when k increases, elements of other classes appear as neighbors, undermining the classification performance.

Although there are six classes being evaluated, the random classifier is able to predict correctly slightly more than 16% (1/6), having an accuracy value around 20%. This happens due to the fact that the topics are not equally distributed in

terms of number of messages, as Figure 9 shows.

6.1.2 Support Vector Machine

Later, a more elaborate classifier, the *Support Vector Machine* (SVM) (LEE *et al.*, 2011), was used. The expectation was that, when a training set was presented to the algorithm, it would identify which users and behaviours were characteristic for each class of message. Thus, when new messages are presented to the model, it is able to identify the patterns learned on the training phase, and classify the message's topic.

For the training process, 80% of the dataset was used for training, producing an input feature vector of dimension 44571×712, and the remaining 20% were separated for testing. The SVM implementation used was the C-Support Vector Classification, available on Python's library Scikit Learn (PEDREGOSA *et al.*, 2011). Different kernels were tested, but the one with best results for the current task was the linear kernel.

From 10 distinct executions of the algorithm, using different training and test samples, a final accuracy of $\mathbf{50.7 \pm 1.7}\%$ was achieved. This result shows that the classifier was able to identify patterns on the training dataset, making the prediction task something feasible, for new samples presented to the model.

Table 2 depicts the confusion matrix of the classification, showing the algorithm's performance for each class. Classes with fewer samples (*world* and *market*) were not classified correctly. The most precise classification involved messages of *politics* and *entertainment*, with over 70% recall. *Sports* messages were mistakenly classified as *everyday* and *entertainment*, what is not very surprising, given the resemblance of these topics.

It also must be noted that the algorithm reached a rate of 100% of accuracy on the training stage, that was not achieved within the test dataset. Attempts to increase the model's generalization capability were conducted, by reducing the training set size. However, although the accuracy of the training stage decreased, the test accuracy

rate did not increased.

For comparison purposes, a matrix T_{random} was created, having the same number of ones' cells (i.e the same density and sum) of matrix T, but with a random distribution. The same algorithm was evaluated on T_{random}.

Interestingly, the SVM was able to find an (imaginary) pattern on the training dataset, also achieving 100% of accuracy on this stage. However, when attempting to predict samples of the test dataset, the model reached $21.5 \pm 4.3\%$ of accuracy, like the KNN's random model.

This comparison shows that, even though the training process may reveal false patterns of behaviour, hindering the learning process, the high accuracy (over 50%) obtained in the test stage for the real dataset is an indicator that there are indeed observable and recurrent features on information diffusion events.

Table 2 – Confusion matrix for classification task, using SVM.
Classes: (a) - everyday; (b) - sports; (c) - world; (d) - politics; (e) - entertainment; (f) - market

		Observed							
		(a)	(b)	(c)	(d)	(e)	(f)	Total	Recall
Expected	(a)	**20**	2	2	3	5	0	32	62.5%
	(b)	5	**15**	1	3	8	0	32	46.8%
	(c)	4	1	**1**	4	7	0	17	5.8%
	(d)	8	4	0	**32**	1	0	45	71.1%
	(e)	2	7	0	1	**28**	1	39	71.8%
	(f)	4	3	1	5	0	**0**	13	0%
	Total	43	32	5	48	49	1	178	53
	Precision	46.5%	46.8%	20.0%	66.7%	57.1%	0%		

6.1.3 Neural Network

Another technique tested was the use of artificial neural networks as classifiers. As with the SVM, neural networks are able to build non-linear functions from the input data (M_i), in order to predict the output class.

The network chosen was the multilayer perceptron (MLP), with hyperbolic tangent as activation function and weights updated using gradient descent and back-propagation. The network's had 44571 ($|M_i|$) input entries and six (one for each class) outputs. The class is than defined by the output with the highest value. 668 samples (75% of the dataset) were presented to the model during the training stage. The remaining 25% samples were used for testing.

Different network configurations were tested and was verified that few neurons were enough to achieve 100% accuracy on the training process (indicating overfitting). Thus, a neural network with only 10 neurons and one hidden layer were enough to achieve good results.

After training, the neural network model was able to correctly classify **52.55%** of the test samples.

Figure 12 show the training and test processes. For each configuration of the network, obtained from the training process, the test dataset is submitted to the network and its accuracy evaluated. Note that after 60 epochs, the training network could classify 100% of the training set, while the best model for unseen test data appeared on epoch 55. Each epoch is characterized by the presentation of all the training samples once.

6.2 Unsupervised Learning

As said before, the class assignment proposed by Folha de São Paulo is only one kind of labeling policy. We can conceive that there are other ways to group messages.

For this, unsupervised clustering algorithm were used to identify new topics from the behaviour vectors M_i. The groups indicated by the algorithms are discovered without human intervention, by the detection of clustering patterns on the database.

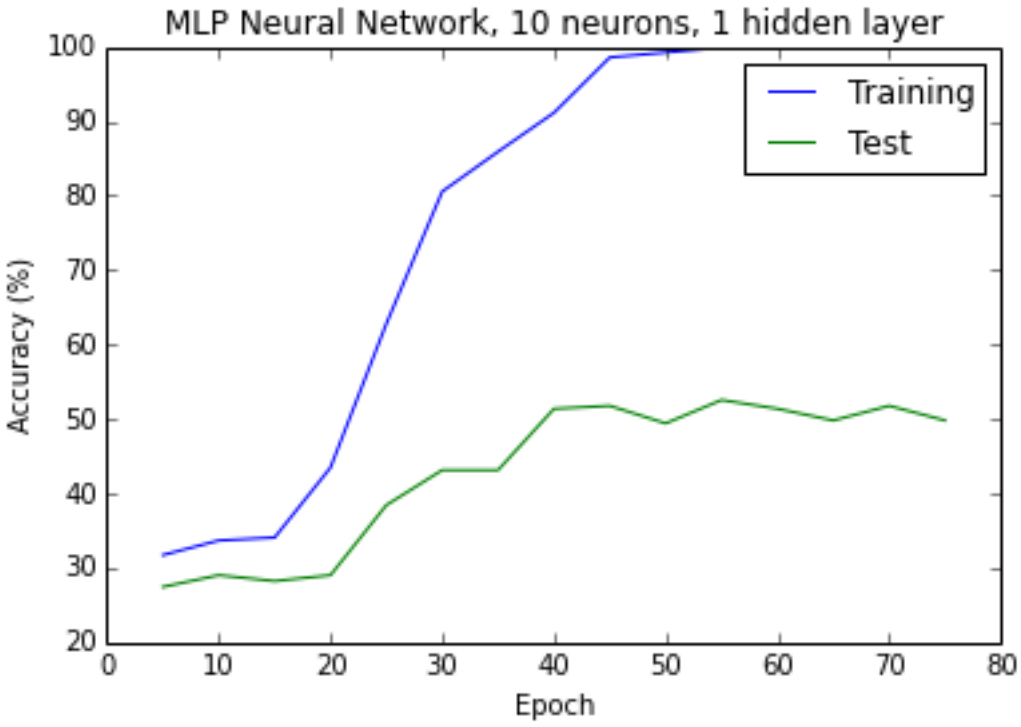


Figure 12 – Multilayer perceptron Neural Network - training and test performance along epochs.

6.2.1 K-Means

With that goal, an attempt was made to cluster messages with near behaviour vectors (M_i) and verify if there was any proximity between their topics. For that, we used a modification of the *k-means* algorithm, using the function cosine as distance metric.

Different group quantities (k) were considered, but the test results were not satisfactory. Firstly, when compared to the standard categories, the obtained clusters had messages with completely different standard categories (like sports and economics together). Besides, in the inspection of the obtained groups, it was not possible to identify clear patterns in the proposed clusters. Table 3 presents some examples of the result, showing five messages that are part of five different detected

clusters.

A possible explanation for this effect is that clusterization is a challenge over high dimensional spaces. As $M_i \in \Re^{44571}$, the data is susceptible to the "curse of dimensionality", so that attempts to group it without a previous preprocessing stage for dimension reduction may not be successful.

6.2.2 LDA

Another clusterization algorithm used was the latent Dirichlet allocation (LDA) (BLEI *et al.*, 2003). This model, generally used on natural language processing, has the promising feature of detecting multi classes in an unsupervised procedure.

The LDA algorithm is generally used for the task of topic detection on documents collections. The algorithm assumes that each document can be associated with a mixture of topics and, based on the frequency of words present in each document and its probability for each topic, returns the document's probability distribution for each topic. The definition of each topic is done based on the general distribution of words (being then unsupervised), but the number of topics sought is predefined by the user.

An adaptation of this algorithm was conducted, in order to use data from social interaction to detect the topic of the messages set M. The tweets (messages) were inputted in the algorithm as documents, but instead of using its words frequence to characterize them, each message was characterized by the users who *retweeted* them. Thus, the LDA model was able to detect the users most representative for each topic and the topics distribution of each message.

The algorithm implementation used was the Python library gensim (REHUREK; SOJKA, 2010), that offers an efficient implementation of LDA and other topic detection algorithms. Distribution for different number of topics were experimented and the results were observed.

Table 4 presents the result for ten topics distribution. For each topic, the most representative tweets (tweets that had the highest probabilities of belonging to

each topic) are shown. Although the topics detected with LDA seem slightly more cohesive than the topics detected with k-means algorithm, it still difficult to affirm that there is a clear separation and subject characterization on the results.

6.3 Attempts of Improvement

It is valuable to note that attempts to improve the (very noisy and sparse) dataset were carried out, in order to achieve better results both in the supervised and in the unsupervised experiments. However, none of the attempts produced experimental improvements.

The first attempt conducted was the reduction of matrix T's dimension. The high number of users involved in diffusions, $U = 44571$, presented a challenge to operations such as distance calculation, as shown in the previous sections.

A viable solution was to remove from U users that did not participate often in messages diffusion. Observation 1 and Figure 7 show that most users had less than 5 *retweets* on the observable database.

Therefore, users with less than 5 *retweets* were removed from U and subsequently from T, resulting in a new matrix with only 826 users and still 890 messages. Nevertheless, the results obtained in all experiments were worse than the ones with the original matrix T, indicating that users with few *retweets* bring relevant information about the content.

Another solution to decrease the number of users, was to group potentially similar users. A community detection algorithm[2] was used to identify groups of users on the network structure G.

Members of U could then be grouped in their respective communities and each community corresponded to a column in T. The value of each cell were evaluated as the total number of *retweets* produced amongst the community.

[2] See Chapter 7 for more information about community detection algorithms.

The algorithm found 2001 communities (many of them small), reducing T to 890×2001. However the experiments conducted with this reduced T did not presented satisfactory results.

Another path taken was the attempt to increase T density, as most of its values are equal to zero.

It is known that, on the observed dataset, the absense of a *retweet* does not necessarily means that a user is not interested in the message being diffused. It can happen that a user is offline during the diffusion or even that his or her friends already *retweeted* a message, being redundant a new sharing of the same content. Therefore, some zeros on T can be considered false negatives and could be replaced by ones.

A method to infer if a user is prone to share a specific message m was conducted from the analysis of its social connections, represented in the graph G. The method consists in verifying how many followees of a user are involved in the diffusion. If message m was *retweeted* by its followees, it can be assumed that the user has a high chance of approving the content being shared.

An exploration to infer if a user is prone to share a specific message m was conducted from the analysis of its social connections, represented in the graph G. If message m was *retweeted* by its followees, it can be assumed that the user has a high chance of approving the content being shared.

Using this principle, an attempt to "propagate" a message's *retweet* was performed, by filling with 1 the cells on T belonging to direct followers of users that *retweeted* a message, thus eliminating possible false negatives.

However, this attempt was also not successful, as the results worsened, indicating that this technique probably raised the false *positive* rate.

6.4 Results Discussion

The results presented in this chapter gives interesting answers to the questions firstly presented in the introduction. The amount of information regarding the subject of a message that can be obtained from observations of user behaviour is surprising. One should note that none of the techniques presented directly analyses the textual content of messages, being all results obtained fruit of network observation.

From the results presented above, the supervised learning seemed to be the most promising. It is fascinating that, even with a very sparse feature space (99,98% of the elements of T are zeros!), it is possible to build classifiers that predict a message's class over six possibilities with an accuracy rate of over 50%. Also, the comparison to random models, shows that the positive obtained results are not result of chance, but in fact product of learning.

The use of such classifiers can be even more promising, if combined with traditional topic detection algorithms that use textual features (natural language processing methods). It is expected that the classifiers based on network features can be able to detect features not present on the textual content, being an option to enhance traditional classifiers, assuming they carry supplementary information.

The evaluation and analysis of unsupervised learning results present an intrinsic challenge, as there are not established "correct answers" that can be used to comparison and benchmark (as what was done to the supervised results). The analysis of pertinence of messages in clusters is a subjective task and therefore debatable. However, even considering this subjective margin, there were not evident results in the analysis done.

The fact that few users that compose the dataset posted more than once (Observation 1), and that there is small coherence on the topics published by them (Observation 2), certainly presented a challenge to the experiments proposed. However, attempts to mitigate it did not seem to be effective, as described in previous section.

It is also important to emphasize that, due to the pioneering nature of this

work, it was not possible to find alternative results on the literature, that could be used for comparison.

Finally, some final remarks can be made, relating directly to the questions presented at the beginning of the chapter:

Observation 4. *It is possible to obtain the direction of a message category by knowing the users who participate in their sharing process.*

Observation 5. *There is a relationship between the category (topic) of a message and the users that share it.*

Table 3 – Examples of clustered tweets using k-means.

Category	Tweet's content
Cluster 1	
esporte	#FolhanaCopa No Twitter, PF (@agenciapf) corneta convocação e ped...
esporte	Rafael Nadal vence Djokovic e conquista o título de Roland Garros...
poder	Alckmin rechaça greve e diz que servidores do Metrô podem ser dem...
mundo	Homem mais velho do mundo morre aos 111 anos em Nova York. http:/...
poder	Metroviários do Rio ameaçam entrar em greve durante a Copa do Mun...
Cluster 2	
mundo	Avião da Malaysia Airlines cai na Ucrânia com 295 pessoas a bordo...
mundo	Uruguai inicia cadastro de quem quer plantar maconha em casa. htt...
esporte	Mayra Aguiar ganha a 1ª medalha de ouro para o Brasil no Mundial ...
esporte	Árbitro relata atos racistas contra Aranha, e Grêmio pode até ser...
esporte	Torcedora que chamou Aranha de "macaco" é afastada do trabalho. h...
Cluster 3	
mundo	Golfinhos militares da Crimeia entram para o exército russo. http...
mundo	Adolescente britânica fica presa em bueiro após tentar pegar celu...
cotidiano	8 em cada 10 brasileiros temem ser torturados pela polícia, diz p...
mercado	Grandes empresas indicam as perguntas que os candidatos a emprego...
esporte	#FolhanaCopa Assim como no Rio, seleção brasileira é recebida com...
Cluster 4	
cotidiano	Presas postam fotos sensuais dentro da cadeia no Paraná. http://t...
mundo	'O que você quer fazer antes de morrer?', diz campanha da Malaysi...
poder	Marina e Aécio querem acabar com o 'Minha Casa', diz Dilma. http:...
entreten.	Cavaleiros do Zodíaco completam 20 anos no Brasil com filme e ret...
esporte	Gremista acusada de racismo pede desculpas e quer encontro com Ar...
Cluster 5	
esporte	Narrador Luciano do Valle morre aos 70 anos. http://t.co/HpEZEjKY...
entreten.	Atriz Fiorella Mattheis leva tiro de borracha na frente de restau...
entreten.	Fotógrafo ferido pela PM nos protestos de 2013 lança mostra em Sã...
poder	Aécio sela união com PMDB no Rio e divide base de Dilma. http://t...
esporte	Para torcida, Fred é o pior jogador do Brasil na Copa, mostra Dat...
Cluster 6	
esporte	"We Are One": com participação do Olodum, clipe oficial da Copa é...
cotidiano	Viga de obra do monotrilho cai e mata uma pessoa em São Paulo. ht...
cotidiano	Nasce no zoo de BH o primeiro filhote de gorila da América do Sul...
cotidiano	Alckmin sanciona nesta sexta-feira lei que proíbe máscaras em pro...
poder	Bandeira eleitoral de Aécio Neves, programa tucano é alvo de inve...

Table 4 – Topics detected using LDA. For each topic, the five tweets most representative are chosen.

Rank	Category	Tweet's content
		Topic 1
1	esporte	Árbitro relata atos racistas contra Aranha, e Grêmio po...
2	esporte	STJ decide que Sport é o único campeão brasileiro de 19...
3	esporte	#FolhanaCopa Júlio César avisou que ia pegar três pênal...
4	esporte	Para torcida, Fred é o pior jogador do Brasil na Copa, ...
5	esporte	Neymar pode ser preso se mostrar a cueca com patrocínio...
		Topic 2
1	entreten.	"Ela é minha heroína": Gata salva menino de 4 anos de a...
2	poder	Justiça de SP barra candidatura de Paulo Maluf a deputa...
3	entreten.	Veja em galeria os 23 convocados de Felipão para a Copa...
4	esporte	#FolhanaCopa Papa pede por uma Copa maravilhosa, disput...
5	esporte	#FolhanaCopa Terceiro melhor jogador do mundo, Ribéry é...
		Topic 3
1	esporte	Maurício de Sousa faz desenho em homenagem a centenário...
2	entreten.	Felicidade de recém-casados se esgota em dois anos, diz...
3	esporte	Casa de gremista que xingou Aranha pega fogo em Porto A...
4	cotidiano	Em greve há quase um mês, professores fazem novo ato e ...
5	entreten.	Foo Fighters confirma quatro shows no Brasil em janeiro...
		Topic 4
1	mundo	Adolescente britânica fica presa em bueiro após tentar ...
2	mercado	Salgaram a cerveja e o refrigerante: governo aumenta o ...
3	cotidiano	Restaurantes da Vila Madalena fecham as portas por falt...
4	cotidiano	Com máxima de 12,6C, cidade de São Paulo registra a ta...
5	entreten.	Foto do dia: Panda gigante come bambu em uma estação ec...
		Topic 5
1	poder	'Não vou me deixar perturbar por ofensas verbais', diz ...
2	esporte	#FolhanaCopa Espetinho custará R$ 15 nos estádios da Co...
3	mundo	Ato lembra os 69 anos da explosão de bomba atômica em H...
4	cotidiano	Vendaval em MT derruba cobertura de aeroporto reformado...
5	cotidiano	Presas postam fotos sensuais dentro da cadeia no Paraná...
		Topic 6
1	poder	Após debate, Dilma defende criminalização da homofobia....
2	esporte	#FolhanaCopa Para evitar vaias, Dilma será 'blindada' n...
3	poder	A dez dias do início da Copa do Mundo, aeroportos estão...
4	poder	Confira o programa de governo de Marina Silva ponto a p...
5	entreten.	Márcio Canuto é derrubado por multidão em SP e sofre 'm...

Continued on next page

Table 4 – Continued from previous page.

Rank	Category	Tweet's content
		Topic 7
1	entreten.	#FolhanaCopa Ao vivo, jornalista Fernanda Gentil chora ...
2	entreten.	Aos 86 anos, Laura Cardoso decreta: 'Ator que não gosta...
3	entreten.	'O Poderoso Chefão' e 'Forrest Gump' voltam a ser exibi...
4	poder	Justiça manda soltar acusados de tráfico em helicóptero...
5	entreten.	Consumo de álcool matou 3,3 milhões de pessoas em 2012,...
		Topic 8
1	esporte	#FolhanaCopa Chilenos tentaram entrar no Maracanã com i...
2	entreten.	Teoria sobre beleza de 'How I Met Your Mother' é compro...
3	entreten.	'Meu balde não tem água, moro em São Paulo', diz Rafinh...
4	entreten.	Hospital ganha prêmio após socorrer cachorro que comeu ...
5	entreten.	Mais um Round da polêmica: aos 40 anos e com crise de i...
		Topic 9
1	cotidiano	Dilma Rousseff é xingada por multidão em festival de ro...
2	entreten.	Bombardeio ilumina o céu sobre a faixa de Gaza durante ...
3	cotidiano	Manifestantes fecham vias e queimam álbum em ato contra...
4	mundo	Foto do dia: Aurora boreal cria 'show de luzes' no Cana...
5	mundo	Teste de DNA inocenta dois negros americanos condenados...
		Topic 10
1	cotidiano	Mulher de 87 anos é atendida no chão em hospital de SP....
2	esporte	O jornal norte-americano "The New York Times" elogiou a...
3	cotidiano	Queda de avião mata ao menos dois em Santos; campanha d...
4	poder	Mulher de Campos pegou voo comercial e está com os filh...
5	entreten.	Eterno Harry Potter, Daniel Radcliffe visita bar de mac...

7 Experiments Set II - Influence of Connections on the Individual Behaviour

This chapter proposes deeper investigations on the network's structure, in order to find new patterns of behaviour and common properties. The fundamental question that can be considered as motivation for the experiments described here is the following:

Fundamental question 2. *Is it possible to determine the behaviour of an individual from his social connections?*

We can affirm that we already have a faithful representation of users social connections, as it was possible to extract information of users relationship from *Twitter*'s service, to form graph G. Also, as done in the previous chapter, we evaluate a user's behaviour from the content published by him/her, in the observed database.

Therefore, to answer Question 2, we can formulate a corresponding practical question:

Practical question 2. *How the connections established by the users relate to the content posted?*

7.1 Community Detection

One way to select groups of individuals that may influence each other is by detecting communities over the network. In a community, there is a concentration of connections among its members and hence an increase on the mutual influence between communities' members.

In order to identify the communities present in network G, a community detection algorithm (ROSVALL; BERGSTROM, 2008) was executed. Several commu-

nities were detected, many with 2 elements (2 connected users, isolated from the rest of the graph), but 74 larger groups with 50 users or more have also been identified.

7.1.1 Messages Frequencies

A first experiment consisted in comparing the frequency at which specific messages were shared inside communities and in the rest of the network.

For each community with a relevant number of users (50 or more), the number of messages *retweeted* for each Folha de São Paulo's original post was computed. A high coherence was verified in the behaviours of individuals belonging to the same community. Table 5 compares the percentage of users that shared a message inside and outside the five largest communities (in number of users). The percentages of the five cases most discrepant for each community are shown, in order.

The cases shown on Table 5 indicate that there is a certain level of coordination in the selection of shared messages by each community. It is interesting to see how some messages are much more emphasized inside a community. Even though they are large communities, Community 1 and 2, for example, present higher rates of message shares than the observed on the whole network.

Equally interesting are cases where the community seems to suppress the spread of a message, as seen in Community 4, where highly *retweeted* messages do not receive the same popularity among its members, compared to its general popularity.

In order to validate this tendency of community coordination, experiments were held with a randomized dataset. For this experiment, the *retweet pattern* of all users were randomly permuted, while the social connections were maintained. Thus, a user i belonging to a community a might now present the behavior of a user j that belongs to a community b.

The five messages with the most discrepant sharing rates are presented on Table 6, for the same communities presented on Table 5. The comparison of the two tables shows that the sharing rates of the randomized dataset are less discrepant than the original dataset. This leads to the conclusion that there are consistent patterns

Table 5 – Comparison of messages sharing rates inside and outside communities. The five biggest communities are presented, with its top five discrepant messages.

Community 4, members: 3579, total tweets: 5791					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	1.0%	1.1%	0.8%	1.0%	0.3%
Sharing rate, outside the community	2.1%	2.1%	1.6%	1.7%	0.9%
Community 1, members: 2345, total tweets: 7353					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	2.9%	2.2%	2.2%	2.1%	2.1%
Sharing rate, outside the community	0.7%	0.2%	0.3%	0.2%	0.3%
Community 2, members: 918, total tweets: 2205					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	5.7%	4.5%	4.1%	4.0%	2.8%
Sharing rate, outside the community	0.3%	0.5%	0.2%	0.4%	0.1%
Community 3, members: 486, total tweets: 923					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	1.6%	0.8%	1.9%	1.0%	1.2%
Sharing rate, outside the community	0.2%	2.0%	0.9%	0.1%	0.3%
Community 22, members: 468, total tweets: 973					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	2.8%	0.2%	2.1%	1.9%	0.6%
Sharing rate, outside the community	0.8%	2.0%	0.6%	0.4%	2.0%

of behaviour inside communities indicating the presence of homophily, that were mischaracterized when the data was shuffled.

7.1.2 Topics Distribution

Another experiment consisted in the analysis of how topics are distributed among communities. For this, the six standard categories were considered and the number of messages published by each community was computed.

First, Figure 13a shows the general distribution of amount of messages per topic for the whole network. Then, Figures 13b to 13g shows the distribution for five

Table 6 – Messages sharing rates on randomized dataset.

Community 4, members: 3579, total tweets: 6782					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	1.3%	0.8%	0.4%	0.2%	0.6%
Sharing rate, outside the community	1.7%	0.5%	0.1%	0.4%	0.4%
Community 1, members: 2345, total tweets: 5022					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	2.5%	1.0%	0.6%	0.8%	0.6%
Sharing rate, outside the community	1.9%	0.5%	0.2%	0.4%	0.2%
Community 2, members: 918, total tweets: 1888					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	2.7%	0.8%	1.3%	0.8%	1.0%
Sharing rate, outside the community	1.6%	0.2%	0.8%	0.2%	0.5%
Community 3, members: 486, total tweets: 940					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	1.4%	1.0%	2.1%	1.4%	1.2%
Sharing rate, outside the community	0.3%	0.1%	1.2%	0.6%	0.4%
Community 22, members: 468, total tweets: 915					
	1st	2nd	3rd	4th	5th
Sharing rate, inside the community	1.3%	1.5%	1.3%	1.5%	1.1%
Sharing rate, outside the community	0.3%	0.6%	0.4%	0.6%	0.2%

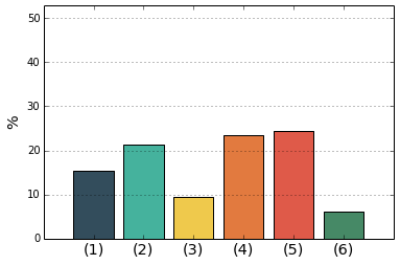
communities, with more than 100 members.

The figures demonstrate how communities' topic distribution may have different profiles from the rest of the network. Community 18, for example, seems to have a high tendence of sharing messages of entertainment and sports, while community 42 is more focused on everyday news.

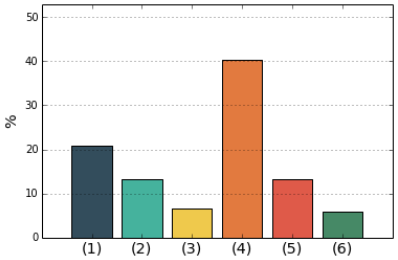
7.1.2.1 TF-IDF

In order to evaluate the importance of individual messages inside communities, a normalization procedure was applied to the database.

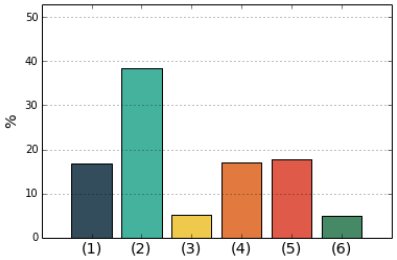
A very common statistic applied generally to collections of textual documents



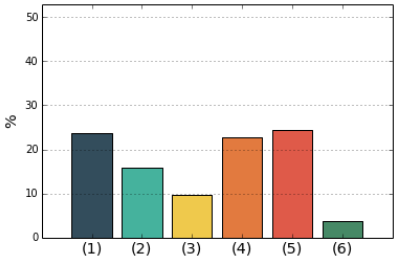
(a) General topic distribution, for the whole network.



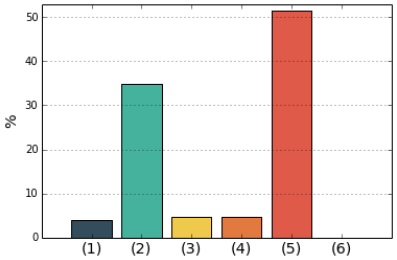
(b) Community 1



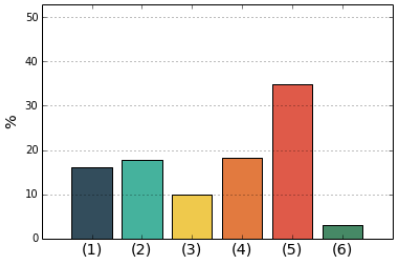
(c) Community 8



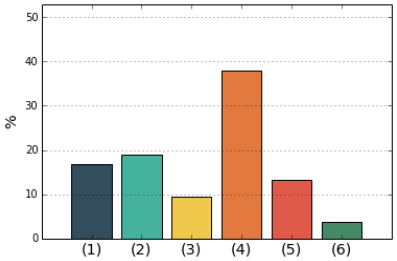
(d) Community 10



(e) Community 17



(f) Community 30



(g) Community 62

Figure 13 – Topic distribution in the entire network and within communities. In the histograms, topics are represented as: (1) “everyday news”; (2) “sports”; (3) “world”; (4) “politics”; (5) “entertainment”; (6) “market”.

is the TF-IDF (DILLON, 1983). In this scenario, the importance of a document's words are evaluated as the product of two indexes: the frequency of the term inside the document (TF) and the inverse of the word's frequency within the documents collection (IDF).

The TF-IDF algorithm was implemented by the author. Different expressions can be used when calculating TF and IDF and the chosen metrics used were:

$$TF(t,d) = \frac{f_{t,d}}{\max f_{t,d}}$$

and

$$IDF(t,D) = \frac{N}{N_t}$$

therefore:

$$score(t,d,D) = TF(t,d) \times IDF(t,D)$$

where d is a document, t a term, D a collection of documents, $f_{t,d}$ the frequencies (counts) of term t in document d, $\max f_{t,d}$ the frequency of the most frequent term in t, N is the number of documents in D and N_t the number of documents that contains t.

From the expressions it is evident that words that are simultaneously frequent inside a document and uncommon in other documents receive a high score and are considered relevant.

This technique can be adapted to the community database. Using the documents collection as analogy, if communities are considered as "documents" (therefore, a group of communities is equivalent to a collection of documents) and tweets as the document's "words", then when TF-IDF statistic is applied to the data, each community's tweet receives a score related to its relevance in the characterization of that community.

The use of this technique on the data resulted in Table 7, where the weighted tweets can be seen for the same six communities of Figure 13. For each community, the five tweets with highest weight are shown, together with its TF-IDF scores.

The analysis of the message's content of each community brings interesting insights about its members. As seen in the data presented, not only the predominant topics can be identified for the communities (Figure 13), but also preferred themes inside those topics.

For example, community 8 seems to have a strong focus on publications related to Brazilian football team Palmeiras, community 17 specializes on pop culture content and 30 is apparently focused on entertainment messages. It is curious that, although the message with highest TF-IDF score of community 30 belongs to class "mundo" (world), its content is in fact related to entertainment. This is one clear evidence that messages can be categorized in different forms (see Section 5.1.1 for more on this).

Going even further, political positions seem to be evident on this characterization. Although the most representative messages of communities 1 and 62 are both about politics, when the messages content is analysed, the first seems to be supportive to the current government, while the second supportive to opposition.

Evidently, this analysis of content cohesion inside communities is a subjective task, similar to the analysis of the results done in Section 6.2. However, the results presented here seem to be much more consistent and explicit than the ones presented in the previous chapter.

7.2 Biclustering

Another way of detecting groups of users is by the use of biclustering algorithms (MADEIRA; OLIVEIRA, 2004). This technique differs from other clustering methods (like the ones explored in Chapter 6) as it considers two dimensions of the data, when detecting groups. Thus, when applied to matrix T, the algorithm finds groups of users that present similar behaviour – i.e. that shared the same group of messages.

The InClose2 algorithm (ANDREWS, 2011) able to find all maximum biclus-

ters on a binary matrix was used. However, as T is extremely sparse, the algorithm did not find large groups, presenting biclusters with a maximum of 6 tweets and 4 users. Moreover, the noisy information present in T (specially false negatives) also affects the size of the found biclusters, as it "breaks" possible patterns.

A test was made to see if there is coherence between the topics posted by members of the same bicluster. To this, the subjects of the messages belonging to the same bicluster were analysed, but no evidence of cohesion on the topics was found, possibly agreeing with Observation 2.

Another experiment conducted was the verification if members of a bicluster had a tendency to be connected on the social network. However, this was not verified on the examined data, leading to the belief that the synchrony of behaviour that characterized the formation of a bicluster was more associated with chance than with any coherence pattern.

7.3 Discussion of the obtained results

The experiments made with the community detection have instigating effects, when it shows that a group made exclusively by exploring connectivity patterns in the social network, without any information regarding content, can also be used to group messages.

Therefore, evidences of the existence of a relationship between social connections and behaviour were shown. However, it is difficult to determine if either the social connections influence the individual behaviour, or if the social connections are a consequence of already consolidated individual behaviours, that may be strengthened when inserted in a social environment characterized by the existence of people with similar beliefs. Nevertheless, the analysis of this relation of cause and effect is out of the scope of this work.

It should be noted that this relation was already explored in works such as McPherson *et al.* (2001), that explores the concept of homophily on networks. Thus,

the ambition of this research is the aggregation of more empirical data to the ongoing literature on the subject, verifying the extent to which homophily is manifested in OSNs, and the advance of computational methods that can be used to analyse other human social structures.

The experiments made with biclustering tried to find a more strict relationship between behaviour and type of message exchange. However, the observed results did not reveal such pattern.

So, to summarize the experiments presented on this chapter, some observations can be outlined:

Observation 6. *There are indications of comportamental coordination within communities.*

Observation 7. *Apparently, some communities can be specialized in topics and this knowledge can be used to characterize group of users and infer about active subjects on the network.*

Table 7 – Most representative tweets according to tf-idf normalization.

tf-idf	Category	Tweet's content
Community 1		
3.98	poder	Comício de Marina reúne 300 pessoas em campo de fut...
3.91	mercado	Produção de petróleo do país cresce quase 15% e bate re...
3.54	cotidiano	Brasil reduziu em 50% o número de pessoas que sofrem fo...
3.24	poder	Aécio se recusa a comentar documento que admite uso...
3.13	cotidiano	Haddad planeja dar desconto no IPTU para empresa com...
Community 8		
4.13	esporte	Seleção brasileira deverá jogar no novo estádio do Palm...
3.61	esporte	Maurício de Sousa faz desenho em homenagem a centenár...
3.40	esporte	WTorre faz homenagem ao centenário do Palmeiras no...
3.39	cotidiano	Polícia apreende quase duas toneladas de maconha em...
3.35	esporte	Palmeiras se reunirá com Dorival Jr. e quer anunciá-lo ...
Community 10		
3.42	cotidiano	Haja nariz: PF prende quadrilha de traficantes e apreen...
3.26	cotidiano	'É preciso ensinar a usar drogas' diz ex-traficante que...
2.96	poder	Promotoria aponta indício de limpeza social em ruas do ...
2.95	mercado	Confiança da indústria cai pela 6ª vez seguida, diz FGV...
2.84	cotidiano	Saída antecipada do trabalho faz lentidão no trânsito d...
Community 17		
3.71	entreten.	Com apenas 22 anos, integrante do One Direction comp...
2.91	entreten.	Faculdade americana oferece aula sobre Miley Cyrus. htt...
2.76	entreten.	Lady Gaga e Céline Dion dão as caras na comédia 'Mupp...
2.71	entreten.	Justin Bieber provoca acidente de carro e é preso novam...
2.67	esporte	Lesão tira David Luiz de amistoso da seleção brasileira...
Community 30		
3.50	mundo	Nintendo diz que não permitirá relações homossexuais em...
2.98	entreten.	Encara? Em quiosque na zona norte de SP, salgados eno...
2.81	entreten.	Casas vendem cachorro-quente com banana e pão de quei...
2.71	mundo	Comandante do Exército tailandês anuncia golpe militar ...
2.69	poder	Jornalista é detida no Rio por filmar prisão de torcedo...
Community 62		
3.76	poder	Ex-diretor da Petrobras cita Lobão, Renan e Henrique Al...
3.42	poder	Demora da Justiça livra Luiz Estevão de duas condenaçõe...
3.31	poder	Alckmin rechaça greve e diz que servidores do Metrô pod...
3.26	entreten.	Felicidade de recém-casados se esgota em dois anos, diz...
3.23	mundo	Hamas anuncia que aceita trégua de 72 horas proposta pe...

8 Conclusion

8.1 Review of Achievements

This work intended to be an initial exploration into the nowadays popular Online Social Networks and into the events that take place on it.

The main motivation for this inquiry was the author's interest in complex phenomena that happen daily on those networks, such as information diffusion processes, collective attention and self-organization. After concluding this stage of the research, it can be said that this work was a good opportunity to deepen in the subject and get more insights on those fascinating processes.

However, more than simply being inspired by the observable existence of such phenomena, the work focused on understanding and even predicting some of the intriguing behaviours of groups of persons using an Online Social Service.

As a byproduct of the research, a survey of computational approaches for online social networks have been conceived, with a summarized version of it being presented at Chapter 2 of this text. As seen on the Chapter, recently, several researchers have been interested on this topic, proposing a wide range of applications. The increasing volume of work produced is an evidence that this is a hot research topic, although it is hard to predict for how long it will stay relevant.

In order to obtain a broad and complete perspective of the problem and also to be able to model and characterize those spatio-temporal events, concepts of the complex systems theory were employed to investigate social networks. We were able to collect, model and analyse real data, extracted from actual online social networking service and verify its properties, using some topological indices of complex networks.

With the use of machine learning algorithms, it was possible to detect patterns on the collected data. Evidences was shown to the fact that it is possible to

have indications of a message's subject from its spatio-temporal features, and that communities of users have cohesion in their behaviour.

8.2 Seeking Answers for the Proposed Questions

The experiments attempted to address the questions proposed in Chapter 4: *"Is it possible to predict behaviour, from external observations of the network?"* (main question).

Although this is a complex question that requires a more advanced of research to be completely answered, the experiments described on Chapters 6 and 7 could capture some interesting aspects of it.

First, the experiments on Chapter 6 showed that it is possible to have a good notion of the subjective aspects of a content being diffused, from the network's characterization. Therefore, the general network characterization seems to be powerful to describe specific aspects of the individuals behaviour, like interests, affinities and the specific content of messages exchanged by him.

Going further in that line, Chapter 7 presented evidences of how social relationships affects the individual behaviour and vice versa. Thus, again, from observing a network structure, it is possible to have relevant information about its components behaviour.

Given this, an initial conclusion to the proposed question may be articulated: based on the experiments, there are concrete evidences that it is possible to predict behaviour, at least partially, from the characterization of the network's properties.

8.3 Future Perspectives

Despite the work's relevance to investigate the proposed questions, the findings presented in this work can also be used to contribute with further research that also uses data from online social networks.

One straightforward contribution, supported by the findings of Chapter 6, is the addition of social features (as the ones used in this work) on current machine learning algorithms for topic detection. As the usual techniques of topic detection use only features of the text when a topic is sought, the complementary use of features extracted from the network is possibly able to bring new and helpful information.

Another potential contribution of this work would be to give support for data intensive research on social sciences. Although the methods used on the experiments are all computational, its empirical discoveries can be used on studies of diverse subjects, such as sociology, psychology and economy.

Other aspects of the current research can still be investigated. Three examples are (a) the investigation of the process of building a network and how it impacts on behaviour, (b) the exploration of the relationship between behaviour and other aspects other than message topics, such as [personal preferences in specific contexts or cultural background, and (c) a discussion on privacy issues and how individual privacy can be affected by the use of network's information.

Bibliography

AARTS, O.; MAANEN, P.-P. van; OUBOTER, T.; SCHRAAGEN, J. M. Online social behavior in twitter: A literature review. In: *2012 IEEE 12th International Conference on Data Mining Workshops*. [S.l.]: IEEE, 2012. p. 739–746. ISBN 978-1-4673-5164-5. Cited on page 11.

ADAMIC, L.; ADAR, E. How to search a social network. *Social Networks*, v. 27, n. 3, p. 187–203, 2005. ISSN 03788733. Cited on page 8.

ADAMIC, L. a.; ADAR, E. Friends and neighbors on the web. *Social Networks*, v. 25, n. 3, p. 211–230, 2003. ISSN 03788733. Cited on page 8.

AJMERA, H. *Latest Social Media users stats, facts and numbers for 2014*. 2014. Disponível em: <http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html>. Cited on page 5.

ALSTOTT, J.; BULLMORE, E.; PLENZ, D. Powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, v. 9, n. 1, p. e85777, 2014. ISSN 1932-6203. Cited on page 46.

AMARAL, L. A. N.; OTTINO, J. M. Complex networks. *The European Physical Journal B - Condensed Matter*, v. 38, n. 2, p. 147–162, 2004. ISSN 1434-6028. Cited on page 15.

ANDREWS, S. In-close2, a high performance formal concept miner. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [S.l.: s.n.], 2011. v. 6828 LNAI, p. 50–62. ISBN 9783642226878. ISSN 03029743. Cited on page 71.

ARDON, S.; BAGCHI, A.; MAHANTI, A.; RUHELA, A.; SETH, A.; TRIPATHY, R. M.; TRIUKOSE, S. 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*. New York, New York, USA: ACM Press, 2013. p. 219–228. ISBN 9781450322638. Cited on page 11.

ASUR, S.; HUBERMAN, B. A. Predicting the future with social media. In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent*

Agent Technology. [S.l.]: IEEE, 2010. v. 1, p. 492–499. ISBN 978-1-4244-8482-9. ISSN 03062619. Cited on page 11.

ASUR, S.; YU, L.; HUBERMAN, B. a. What trends in chinese social media. *SSRN Electronic Journal*, 2011. ISSN 1556-5068. Cited on page 8.

BAO, P.; SHEN, H.-W.; HUANG, J.; CHENG, X. Popularity prediction in microblogging network: A case study on sina weibo. *arXiv preprint arXiv:1304.4324*, International World Wide Web Conferences Steering Committee, p. 2–3, 2013. Cited on page 8.

BARABáSI, A.-L. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, 1999. ISSN 00368075. Cited 3 times on pages 9, 24, and 25.

BARBOSA, L. M.; ATTUX, R.; GODOY, A. Multiplex network approach for scientific articles. *NetSci-x2015 - Network Science Conference*, Rio de Janeiro, Brazil, 2015. Cited on page 30.

BARRETT, L.; HENZI, P.; RENDALL, D. Social brains, simple minds: does social complexity really require cognitive complexity? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, The Royal Society, v. 362, n. 1480, p. 561–75, 2007. ISSN 0962-8436. Cited on page 17.

BENEVENUTO, F.; RODRIGUES, T.; CHA, M.; ALMEIDA, V. Characterizing user behavior in online social networks. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference - IMC '09*. New York, New York, USA: ACM Press, 2009. p. 49. ISBN 9781605587714. ISSN 1605587710. Cited on page 5.

BISHOP, C. M.; NASRABADI, N. M. *Pattern recognition and machine learning*. [S.l.]: springer New York, 2006. v. 1. Cited on page 37.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, JMLR.org, v. 3, p. 993–1022, 2003. ISSN 1532-4435. Cited on page 56.

BOLLEN, J.; GONCALVES, B.; RUAN, G.; MAO, H. Happiness is assortative in online social networks. *Artificial life*, v. 17, n. 3, p. 237–251, 2011. ISSN 1064-5462. Cited 2 times on pages 11 and 28.

BOLLOBáS, B.; RIORDAN, O. The diameter of a scale-free random graph. *Combinatorica*, Springer, Secaucus, NJ, EUA, v. 24, n. 1, p. 5–34, 2004. ISSN 0209-9683. Cited on page 26.

BORGE-HOLTHOEFER, J.; nOS, R. a. B.; GONZáLEZ-BAILóN, S.; MORENO, Y. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, v. 1, n. 1, p. 3–24, 2013. ISSN 2051-1310. Cited on page 11.

BORGE-HOLTHOEFER, J.; RIVERO, A.; MORENO, Y. Locating privileged spreaders on an online social network. *Physical Review E*, v. 85, n. 6, p. 066123, 2012. ISSN 1539-3755. Cited 2 times on pages 11 and 12.

BOYD, D. *Big Data: Opportunities for Computational and Social Sciences*. 2010. Http://www.zephoria.org/thoughts/archives/2010/04/17/big-data-opportunities-for-computational-and-social-sciences.html. Disponível em: <http://www.zephoria.org/thoughts/archives/2010/04/17/big-data-opportunities-for-computational-and-social-sciences.html>. Cited on page 10.

CASTILLO, C.; MENDOZA, M.; POBLETE, B. Information credibility on twitter. In: ACM. *Proceedings of the 20th international conference on World wide web - WWW '11*. New York, New York, USA: ACM Press, 2011. p. 675. ISBN 9781450306324. Cited on page 12.

CHA, M.; HADDAI, H.; BENEVENUTO, F.; GUMMADI, K. P. Measuring user influence in twitter : The million follower fallacy. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, v. 10, p. 10–17, 2010. Cited on page 12.

CHA, M.; MISLOVE, A.; GUMMADI, K. P. 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*. New York, New York, USA: ACM Press, 2009. p. 721. ISBN 9781605584874. Cited on page 8.

CHEN, J.; NAIRN, R.; NELSON, L.; BERNSTEIN, M.; CHI, E. Short and tweet. In: *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. New York, New York, USA: ACM Press, 2010. p. 1185. ISBN 9781605589299. Cited on page 11.

CHEONG, M.; RAY, S. A literature review of recent microblogging developments. *Victoria, Australia: Clayton School of Information Technology, Monash University.*, 2011. Cited on page 7.

CIOFFI-REVILLA, C. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, v. 2, n. 3, p. 259–271, 2010. ISSN 19395108. Cited on page 10.

CONTE, R.; GILBERT, N.; BONELLI, G.; CIOFFI-REVILLA, C.; DEFFUANT, G.; KERTESZ, J.; LORETO, V.; MOAT, S.; NADAL, J. P.; SANCHEZ, A.; NOWAK, A.; FLACHE, A.; San Miguel, M.; HELBING, D. Manifesto of computational social science. *The European Physical Journal Special Topics*, v. 214, n. 1, p. 325–346, 2012. ISSN 1951-6355. Cited on page 10.

COSTA, L. F.; OLIVEIRA, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; Villas Boas, P. R.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. da. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Advances in Physics*, v. 60, n. 3, p. 103, 2007. ISSN 0001-8732. Cited on page 9.

COSTA, L. F. D. F.; RODRIGUES, F. a.; TRAVIESO, G.; Villas Boas, P. R. Characterization of complex networks: A survey of measurements. *Advances in Physics*, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007. ISSN 0001-8732. Cited 2 times on pages 18 and 28.

De Longueville, B.; SMITH, R. S.; LURASCHI, G. "omg, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In: *Proceedings of the 2009 International Workshop on Location Based Social Networks - LBSN '09*. New York, New York, USA: ACM Press, 2009. p. 73. ISBN 9781605588605. Cited on page 11.

DE SOLA POOL, I.; KOCHEN, M. Contacts and influence. *Social Networks*, Elsevier, v. 1, n. 1, p. 5–51, 1978. ISSN 03788733. Cited 2 times on pages 1 and 5.

DESCARTES, R. *A Discourse on the Method*. [S.l.: s.n.], 1637. ISBN 0192825143. Cited on page 15.

DILLON, M. Introduction to modern information retrieval. *Information Processing & Management*, McGraw-Hill, Inc., New York, NY, USA, v. 19, n. 6, p. 402–403, 1983. ISSN 03064573. Cited on page 70.

DOW, P. A.; FRIGGERI, A. The anatomy of large facebook cascades. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, p. 145–154, 2013. Cited on page 7.

DRAIEF, M.; MASSOULI, L. *Epidemics and Rumours in Complex Networks*. [S.l.]: Cambridge University Press, 2010. ISBN 0521734436, 9780521734431. Cited on page 11.

ERDöS, P.; RéNYI, A. On random graphs. *Publicationes Mathematicae Debrecen*, v. 6, p. 290–297, 1959. Cited 2 times on pages 23 and 24.

FACEBOOK. *Company Info | Facebook Newsroom*. 2014. Disponível em: <http://newsroom.fb.com/company-info/>. Cited on page 7.

GAMAL, A. E.; KIM, Y.-H.; El Gamal, A. *Network information theory*. [S.l.]: Cambridge University Press, 2011. Cited on page 2.

GAO, Q.; ABEL, F.; HOUBEN, G. J.; YU, Y. A comparative study of users' microblogging behavior on sina weibo and twitter. In: MASTHOFF, J.; MOBASHER, B.; DESMARAIS, M. C.; NKAMBOU, R. (Ed.). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Berlin, Heidelberg: Springer, 2012. v. 7379 LNCS, p. 88–101. ISBN 9783642314537. ISSN 03029743. Cited on page 8.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, v. 99, n. 12, p. 7821–6, 2002. ISSN 0027-8424. Cited on page 24.

GODOY, A.; VON ZUBEN, F. J. Topology of social networks and efficiency of collective intelligence methods. In: *Proceeding of the Fifteenth Genetic and Evolutionary Computation Conference - GECCO '13*. New York, New York, USA: ACM Press, 2013. p. 1415–1422. ISBN 9781450319645. Cited on page 2.

GOMEZ-GARDENES, J.; REINARES, I.; ARENAS, A.; FLORÍA, L. M. Evolution of cooperation in multiplex networks. *Scientific reports*, Nature Publishing Group, v. 2, p. 620, 2012. ISSN 2045-2322. Cited on page 30.

GÓMEZ, S.; DÍAZ-GUILERA, A.; GÓMEZ-GARDEÑES, J.; PÉREZ-VICENTE, C. J.; MORENO, Y.; ARENAS, A. Diffusion dynamics on multiplex networks. *Physical Review Letters*, v. 110, n. 2, p. 028701, 2013. ISSN 0031-9007. Cited on page 30.

GONZALEZ-BAILON, S.; BORGE-HOLTHOEFER, J.; MORENO, Y. Broadcasters and hidden influentials in online protest diffusion. *American Behavioral Scientist*, v. 57, n. 7, p. 943–965, 2013. ISSN 0002-7642. Cited on page 12.

GRUHL, D.; GUHA, R.; KUMAR, R.; NOVAK, J.; TOMKINS, A. The predictive power of online chatter. In: *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*. New York, New York, USA: ACM Press, 2005. p. 78. ISBN 159593135X. Cited on page 11.

GRUHL, D.; LIBEN-NOWELL, D.; GUHA, R.; TOMKINS, A. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, ACM Press,

New York, New York, USA, v. 6, n. 2, p. 43–52, 2004. ISSN 19310145. Cited on page 8.

GUO, Z.; LI, Z.; TU, H. Sina microblog: An information-driven online social network. In: *2011 International Conference on Cyberworlds*. [S.l.]: IEEE, 2011. p. 160–167. ISBN 978-1-4577-1453-5. Cited on page 8.

GUPTA, M.; ZHAO, P.; HAN, J. Evaluating event credibility on twitter. In: CITESEER. *SIAM International Conference on Data Mining*. [S.l.], 2012. p. 153–164. Cited on page 12.

HALU, A.; ZHAO, K.; BARONCHELLI, A.; BIANCONI, G. Connect and win: The role of social networks in political elections. *EPL (Europhysics Letters)*, v. 102, n. 1, p. 16002, 2013. ISSN 0295-5075. Cited on page 2.

HOGG, T.; LERMAN, K. Stochastic models of user-contributory web sites. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, ACM Press, New York, New York, USA, p. 50–57, 2009. Cited on page 8.

HOLT, R. *Twitter in numbers*. 2013. Disponível em: <http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html>. Cited on page 1.

HUTCHINS, E. Distributed cognition. *Internacional Enciclopedia of the Social and Behavioral Sciences*, IESBS, p. 1–10, 2000. Cited on page 17.

KAISLER, S.; ARMOUR, F.; ESPINOSA, J. A.; MONEY, W. Big data: Issues and challenges moving forward. *46th Hawaii International Conference on System Sciences (HICSS)*, IEEE, p. 995–1004, 2013. ISSN 1530-1605. Cited on page 41.

KARINTHY, F. Chain-links. *Everything is the Other Way*, 1929. Cited on page 25.

KUHN, T. S. *The structure of scientific revolutions*. [S.l.]: University of Chicago press, 1962. 264 p. ISBN 9780226458113. Cited on page 15.

KUMAR, R.; NOVAK, J.; TOMKINS, A. Structure and evolution of online social networks. In: YU, P. S.; HAN, J.; FALOUTSOS, C. (Ed.). *Link Mining: Models, Algorithms, and Applications*. New York, NY: Springer New York, 2010. p. 337–357. ISBN 978-1-4419-6514-1, 978-1-4419-6515-8. Cited on page 8.

KUMAR, S. Analyzing the facebook workload. In: *2012 IEEE International Symposium on Workload Characterization (IISWC)*. [S.l.]: IEEE, 2012. p. 111–112. ISBN 978-1-4673-4532-3. Cited on page 7.

KURKA, D. B.; GODOY, A.; Von Zuben, F. J. Online social network analysis - a survey of research applications in computer science. *Preprint*, 2015. Cited 2 times on pages 5 and 10.

KWAK, H.; LEE, C.; PARK, H.; MOON, S. What is twitter, a social network or a news media? In: *Proceedings of the 19th international conference on World wide web - WWW '10*. New York, New York, USA: ACM Press, 2010. p. 591. ISBN 9781605587998. Cited on page 11.

LANSDALL-WELFARE, T.; LAMPOS, V.; CRISTIANINI, N. 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*. New York, New York, USA: ACM Press, 2012. p. 1221. ISBN 9781450312301. Cited on page 11.

LAZER, D.; PENTLAND, A.; ADAMIC, L.; ARAL, S.; BARABASI, A.-L.; BREWER, D.; CHRISTAKIS, N.; CONTRACTOR, N.; FOWLER, J.; GUTMANN, M.; JEBARA, T.; KING, G.; MACY, M.; ROY, D.; Van Alstyne, M. Social science. computational social science. *Science (New York, N.Y.)*, v. 323, n. 5915, p. 721–3, 2009. ISSN 1095-9203. Cited on page 10.

LEE, C.-H.; CHIEN, T.-F.; YANG, H.-C. An automatic topic ranking approach for event detection on microblogging messages. In: *2011 IEEE International Conference on Systems, Man, and Cybernetics*. [S.l.]: IEEE, 2011. p. 1358–1363. ISBN 978-1-4577-0653-0. Cited on page 52.

LEE, K.; MAHMUD, J.; CHEN, J.; ZHOU, M.; NICHOLS, J. 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*. New York, New York, USA: ACM Press, 2014. p. 247—-256. ISBN 9781450321846. Cited on page 11.

LERMAN, K.; GHOSH, R. 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)*, p. 90–97, 2010. ISSN 00846570. Cited on page 11.

MADEIRA, S. C.; OLIVEIRA, A. L. *Biclustering algorithms for biological data analysis: A survey*. IEEE, 2004. 24–45 p. Disponível em: <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1324618>. Cited on page 71.

MARCHIORI, M.; LATORA, V. Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, v. 285, n. 3-4, p. 539–546, 2000. ISSN 03784371. Cited on page 21.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, Annual Reviews, v. 27, n. 1, p. 415–444, 2001. ISSN 0360-0572. Cited 2 times on pages 28 and 72.

MENDOZA, M.; POBLETE, B.; CASTILLO, C. Twitter under crisis. In: *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*. New York, New York, USA: ACM Press, 2010. p. 71–79. ISBN 9781450302173. Cited on page 12.

MILGRAM, S. The small world problem. *Psychology today*, v. 1, n. 1, p. 61–67, 1967. Cited on page 25.

MILO, R.; SHEN-ORR, S.; ITZKOVITZ, S.; KASHTAN, N.; CHKLOVSKII, D.; ALON, U. Network motifs: simple building blocks of complex networks. *Science (New York, N.Y.)*, v. 298, n. 5594, p. 824–7, 2002. ISSN 1095-9203. Cited on page 27.

MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC '07*. New York, New York, USA: ACM Press, 2007. p. 29. ISBN 9781595939081. ISSN 09538984. Cited on page 8.

MITCHELL, M. *Complexity: A Guided Tour*. [S.l.]: Oxford University Press, 2009. 368 p. Cited 5 times on pages 2, 9, 15, 20, and 24.

MITCHELL, T. M. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, 1997. Cited on page 37.

MUCHA, P. J.; RICHARDSON, T.; MACON, K.; PORTER, M. a.; ONNELA, J.-P. Community structure in time-dependent, multiscale, and multiplex networks. *Science (New York, N.Y.)*, v. 328, n. 5980, p. 876–8, 2010. ISSN 1095-9203. Cited on page 30.

ONG, J. *China's Sina Weibo Grew 73% in 2012 to 500M Accounts*. 2013. Disponível em: <http://thenextweb.com/asia/2013/02/21/chinas-sina-weibo-grew-73-in-2012-passing-500-million-registered-accounts/>. Cited on page 8.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking:bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, 1998. Cited on page 23.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Cited on page 52.

QU, Y.; HUANG, C.; ZHANG, P.; ZHANG, J. Microblogging after a major disaster in china. In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*. New York, New York, USA: ACM Press, 2011. p. 25. ISBN 9781450305563. Cited on page 8.

REHUREK, R.; SOJKA, P. Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010. p. 45–50. Cited on page 56.

RICE, S. a. The identification of blocs in small political bodies. *The American Political Science Review*, Cambridge University Press, v. 21, n. 3, p. 619, 1927. ISSN 00030554. Cited on page 5.

ROGERS, R. Debanalizing twitter. In: *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*. New York, New York, USA: ACM Press, 2013. p. 356–365. ISBN 9781450318891. Cited 2 times on pages 7 and 8.

ROMERO, D. M.; MEEDER, B.; KLEINBERG, J. Differences in the mechanics of information diffusion across topics. In: *Proceedings of the 20th international conference on World wide web - WWW '11*. New York, New York, USA: ACM Press, 2011. p. 695. ISBN 9781450306324. ISSN 14602059. Cited on page 11.

ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, v. 105, n. 4, p. 1118–23, 2008. ISSN 1091-6490. Cited on page 65.

SALATHÉ, M.; VU, D. Q.; KHANDELWAL, S.; HUNTER, D. R. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science*, Springer, v. 2, n. 1, p. 4, 2013. ISSN 2193-1127. Cited on page 12.

SARCEVIC, A.; PALEN, L.; WHITE, J.; STARBIRD, K.; BAGDOURI, M.; ANDERSON, K. "beacons of hope" in decentralized coordination. In: ACM. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. New York, New York, USA: ACM Press, 2012. p. 47. ISBN 9781450310864. Cited 2 times on pages 1 and 12.

SHEKHAR, S.; OLIVER, D. Computational modeling of spatio-temporal social networks: A time-aggregated graph approach. In: *Specialist Meeting-Spatio-Temporal Constraints on Social Networks*. [S.l.: s.n.], 2010. p. 6–10. Cited on page 11.

SHEN-ORR, S. S.; MILO, R.; MANGAN, S.; ALON, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, Nature Publishing Group, v. 31, n. 1, p. 64–8, 2002. ISSN 1061-4036. Cited on page 27.

STARBIRD, K.; PALEN, L. (how) will the revolution be retweeted? In: ACM. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. New York, New York, USA: ACM Press, 2012. p. 7. ISBN 9781450310864. Cited on page 12.

STIEGLITZ, S.; DANG-XUAN, L. 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*. [S.l.]: IEEE, 2012. p. 3500–3509. ISBN 978-1-4577-1925-7. ISSN 15301605. Cited on page 11.

SUH, B.; HONG, L.; PIROLLI, P.; CHI, E. H. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: *2010 IEEE Second International Conference on Social Computing*. [S.l.]: IEEE, 2010. p. 177–184. ISBN 978-1-4244-8439-3. ISSN 1942597X. Cited on page 11.

SUN, E.; ROSENN, I.; MARLOW, C. a.; LENTO, T. M. 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)*. [S.l.: s.n.], 2009. p. 146–153. ISBN 978-1-57735-421-5. Cited on page 7.

SUN, Y.; HAN, J. Mining heterogeneous information networks: Principles and methodologies. *Morgan & Claypool Publishers*, Morgan & Claypool Publishers, v. 3, n. 2, p. 1–159, 2012. ISSN 2151-0067. Cited on page 30.

TRIPATHY, R. M.; BAGCHI, A.; MEHTA, S. A study of rumor control strategies on social networks. In: ACM. *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*. New York, New York, USA: ACM Press, 2010. p. 1817. ISBN 9781450300995. Cited on page 12.

TUMASJAN, A.; SPRENGER, T.; SANDNER, P.; WELPE, I. Predicting elections with twitter: What 140 characters reveal about political sentiment. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, v. 10, p. 178–185, 2010. ISSN 00219258. Cited on page 11.

TYLER, J. R.; WILKINSON, D. M.; HUBERMAN, B. a. E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, v. 21, n. 2, p. 143–153, 2005. ISSN 0197-2243. Cited on page 8.

VIEWEG, S.; HUGHES, A. L.; STARBIRD, K.; PALEN, L. Microblogging during two natural hazards events. In: *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. New York, New York, USA: ACM Press, 2010. p. 1079. ISBN 9781605589299. Cited on page 12.

WATTS, D. J.; DODDS, P. S. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, v. 34, n. 4, p. 441–458, 2007. ISSN 0093-5301. Cited on page 2.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature*, Macmillan Magazines Ltd., v. 393, n. 6684, p. 440–2, 1998. ISSN 0028-0836. Cited 3 times on pages 9, 24, and 26.

WU, F.; HUBERMAN, B. a. Novelty and collective attention. *Proceedings of the National Academy of Sciences of the United States of America*, v. 104, n. 45, p. 17599–601, 2007. ISSN 0027-8424. Cited on page 8.

YANG, F.; LIU, Y.; YU, X.; YANG, M. Automatic detection of rumor on sina weibo. In: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12*. New York, New York, USA: ACM Press, 2012. v. 2, p. 1–7. ISBN 9781450315463. Cited on page 8.

ZASLAVER, A.; MAYO, A. E.; ROSENBERG, R.; BASHKIN, P.; SBERRO, H.; TSALYUK, M.; SURETTE, M. G.; ALON, U. Just-in-time transcription program in metabolic pathways. *Nature genetics*, v. 36, n. 5, p. 486–91, 2004. ISSN 1061-4036. Cited on page 27.