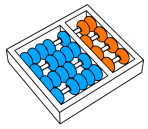
Universidade Estadual de Campinas
Instituto de Computação



Felipe Maciel Cardoso

Topical Homophily in an Online Social Network

Homofilia por Tópicos em uma Rede Social Online

CAMPINAS
2016

Felipe Maciel Cardoso

Topical Homophily in an Online Social Network

Homofilia por Tópicos em uma Rede Social Online

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. André Santanchè

Este exemplar corresponde à versão final da Dissertação defendida por Felipe Maciel Cardoso e orientada pelo Prof. Dr. André Santanchè.

CAMPINAS
2016

Agência(s) de fomento e nº(s) de processo(s): CAPES, 1629113; CAPES, 1461658

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

C179t

Cardoso, Felipe Maciel, 1988-
Topical homophily in an online social network / Felipe Maciel Cardoso. – Campinas, SP : [s.n.], 2016.

Orientador: André Santanchè.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Redes sociais. 2. Twitter. 3. Ciência social computacional. 4. Redes complexas. I. Santanchè, André,1968-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Homofilia por tópicos em uma rede social online
Palavras-chave em inglês:
Social networks
Twitter
Computational social science
Complex networks
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
André Santanchè [Orientador]
Francisco Aparecido Rodrigues
Ricardo da Silva Torres
Data de defesa: 01-08-2016
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Felipe Maciel Cardoso

Topical Homophily in an Online Social Network

Homofilia por Tópicos em uma Rede Social Online

Banca Examinadora:

- Prof. Dr. André Santanchè
  IC UNICAMP
- Prof. Dr. Francisco Aparecido Rodrigues
  ICMC/USP
- Prof. Dr. Ricardo da Silva Torres
  IC/UNICAMP

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 01 de agosto de 2016

Believe those who are seeking the truth.
Doubt those who find it.

André Gide

Acknowledgements

It is impossible for me to thank everyone that contributed to this work. Every article read, coffee talk, discussion and comment helped me with ideas. However, some people were essential and I can not finish this text without mentioning them.

André, a passionate professor and the advisor everyone wishes to have. Thank you for accepting to advise a student that could be more distracted than you. Your love for research and teaching has influenced me greatly. Thank you for the hours of nice talk and the support you have always shown. Your belief in your students should never be undervalued.

Amanda, thanks for your love, patience, and kindness. This period would be a lot harder to endure without your support. You have always motivated me to pursue my goals and your presence is the reason for most of my smiles.

My mother Zu and my brother Marcelo. You have always been my safe harbor and I have always known that I could count on you. I would never be where I am now without your help and advice.

My deepest gratitude for all of my family and friends. You are extremely important in my life. Galizé, the gentlest friend that someone can have. Livia, Bichu, Gabriel, Ana, Jan, Vitor, Marina, Bob, Randerson, Tais, Gra, William and every other lunatic friend from Muchagrama, you are responsible for my most joyful and unexpected moments. Everyone from Cana Sutra, you provided me the most incredible memories. All my friends from climbing, you have influenced me greatly and gifted me with travels to the most spectacular places. All friends from LIS, always available for a nice coffee and conversation, your insights and critiques were fundamental for my growth as a researcher. Burgo, Fabi, Serginho, and Elisa, thank you for your hospitality in my last days in Campinas.

The Pe and López families and every friend from Spain, thank you for the kindness and hospitality. You made these months a lot easier and delightful.

Prof. Yamir Moreno and everyone from COSNET, thank you for the lessons and for giving me a direction in this research. Guilherme, was nice to have an equally adrift friend in Spain, thank you for your companionship and advice.

Finally, I would like to thank CAPES, Microsoft, and Santander[1] for supporting me during my master's degree and the Institute of Computing staff that helped me during these years.

[1]Work also financed by FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and individual grants from CNPq. The opinions expressed in this work do not necessarily reflect those of the funding agencies.

Resumo

Entender a dinâmica de interações sociais é crucial para o entendimento questões que envolvem o comportamento humano. O surgimento de mídias sociais online, tal como Facebook e Twitter, possibilitou o acesso a dados de relacionamentos de pessoas em larga escala. Essas redes são orientadas à informação, com seus usuários compartilhando e consumindo informação. Nesta dissertação, estamos interessados na presença de homofilia por tópicos em uma rede social. Especificamente, nós exploramos como as conexões entre indivíduos estão ligadas com a sua similaridade por tópicos, i.e., a sua proximidade em consideração com os diferentes tipos de conteúdo que circulam pela rede. Para fazê-lo, representamos usuários utilizando as informações de suas mensagens. Nossos resultados demonstram que usuários, na média, estão conectados com usuários similares a eles e que interações mais fortes estão relacionadas com uma alta similaridade por tópicos. Nós também verificamos que, quando se considera apenas usuários dentro de um tópico, a homofilia se manifesta diferentemente de acordo com o tópico. Nós acreditamos que esta pesquisa, além de fornecer uma maneira de aferir similaridade por tópicos, aumenta as evidências de homofilia entre indivíduos, contribuindo para um melhor entendimento de como sistemas sociais complexos são estruturados.

Abstract

Understanding the dynamics of social interactions is crucial to address questions involving human behavior. The emergence of online social medias, such as Facebook and Twitter, has enabled the access to data of people relationships at a large scale. These networks are information oriented, with users sharing and consuming information. In this dissertation, we are interested in the presence of topical homophily in an online social network. Specifically, we explore how individuals connections are related to their topical similarity, i.e., their proximity regarding the different kinds of content that are shared in the network. To do so, we represent users using the information of their messages. Our results show that users, on average, are connected with users which are similar to them and that stronger interactions are related to a high topical similarity. We also verified that, when considering only users inside a topic, homophily manifests differently according to the topic. We believe that this research, besides providing a way to assess the topical similarity of users, deepens the evidence of homophily among individuals, contributing to a better understanding of how complex social systems are structured.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this dissertation, we present our research, which uses data from Twitter to explore the behavior of individuals connections. The emergence of this kind of online social networking service allows testing some social hypothesis with a massive amount of data, giving some insights that would be overwhelming to be obtained by traditional approaches [63]. Furthermore, although the connections in an online social network are only the ties of a specific environment, they might provide data to deepen the understanding of how social systems are structured. Examples can be seen in the study of protests recruitment [30], the limit in the number of friends that someone can have [29], and how individuals tend to acquire the behavior of people they interact with [23].

Moreover, Twitter is often categorized as an information network [52], i.e., it is often a medium for the consumption and sharing of content, which is diffused through users connections. Users decide to follow others, subscribing to receive their posts, i.e., their tweets and retweets. Users can also mention each other, which is another type of connection. This constitution conducts to an interesting linkage between information and connections among individuals, which is the focus of our investigation in this work.

The original purpose of this research was to explore how the evolution of the Twitter social network – i.e., its dynamics of nodes and edges – was related to the flow of information. In this case, nodes and edges stand for users and their *follow* ties, respectively. The information flow is captured by their messages traversing the network through their connections. Some previous works provided evidence that users are influenced by the diffused content [76], but we were interested in how some patterns of information diffusion could affect the links dynamics [54]. We had strong limitations to address this problem since it needed a large longitudinal data set of the social network, including *follow* connections and messages sharing data, which have shown infeasible of being obtained through Twitter's public API. Therefore, we changed the focus to the research presented in this dissertation, exploring a similar question also concerning users connections and the information shared by them. However, it is not concerned with the flow of information, whereas with the user's affiliation in different topics of information. Topics of information, here, stands for the sets of messages that have a semantic association and users associate themselves with topics in different ways [74].

This question is explored through the lens of the homophily principle, which is the tendency of individuals to establish ties with alike [49, 48, 55, 38]. Homophily specifies that

individuals are prone to be connected according to the characteristics they have in common and its presence has been evidenced with respect to sociodemographic characteristics such as race, age, religion, and gender [49]. Nevertheless, in this work we propose an analysis of topical homophily, which is the tendency of individuals to be connected with alike according to their topical similarity. We define their topical similarity according to their affiliation in topics of information, i.e., according to topics that they adopted while sharing messages. Thereby, in this work we provide a method to assess users topical similarity in Twitter, detecting topics of information in a similar fashion to Weng and Menczer [74].

In this work, we want to assess how the information shared by users can be related to the connections of the social network. We begin by verifying the topical homophily in the network and the different degrees that it appears. Our goal is to assure its existence and that it can be related to stronger relationships and more interactions. Furthermore, we also test other complex hypothesis involving the topics of information and user relationships. We are able to give sound answers for some of them, however, they are left open for future work. We further give an overview of our main results.

We show evidence of homophily among users both while following – the act of subscribing to receive another user messages – and mentioning – explicitly mentioning another user in a message. Our results show that, on Twitter, users connected are more likely to have a higher topical similarity than a random pair of users. Furthermore, we verified that *mentions* and *follow* relationships tend to have a similar homophily pattern, despite the belief that they are relationships of a different kind [31]. We also verified that connections with strong interactions tend to be more homophilic. Our analysis could show that the information shared by users could foster the prediction of their connections. We found that users which have a high average similarity with their friends are predominantly connected with the users most similar to them. This was achieved by a proposed mechanism to predict user connections indicating that, for some users, most of the connections are as high similar as possible in the network.

We also assessed the probabilities of users following and mentioning another if they shared topics of information. Our results show that the majority of users tend to establish relationships with users that share some topic and, when they share the topic whereof the user has posted more tweets, the probabilities are significantly higher. Furthermore, we show that different groups of users affiliated with different topics tend to mention and have different levels of topical homophily.

This project is a result of a collaboration with the Spanish laboratory COSNET[1], which started in a six-month exchange program done by the student. All of our data were captured by the Twitter public API. We intend to make this data and the developed code public as soon as this work is published.

The remaining of the document is organized as follows. Chapter 2 presents a bibliographic review of the area and related works; Chapter 3 summarizes the Twitter characteristics that are relevant to us and outlines the process of data gathering; Chapter 4 describes our model to obtain the topics of information; Chapter 5 shows our results regarding user topical homophily; Chapter 6 explores the particularities of users inside topics with respect to their similarities and their probabilities of establishing relationships;

[1] http://cosnet.bifi.es/

finally, Chapter 7 discourses on the results obtained by the analyses.

Chapter 2

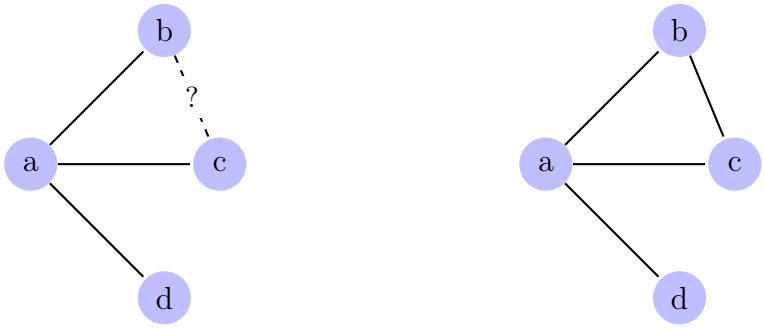
Foundations and Related Work

This dissertation comprehends a social network analysis research and its core involves investigating the relations between the information that individuals in a network share and the structure of connections among them. Our approach is based on the homophily hypothesis that individuals tend to be connected with alike. In this work, we are interested in the information shared and consumed by users to define their similarity. In this chapter, we summarize the main related works in the area and the ones that are directly related to this dissertation. Firstly, Section 2.1 introduces works that study the structural changes of networks over time; Section 2.2 presents works concerned with information on social networks; Section 2.3 presents the concept and the works related to homophily.

This chapter includes the presentation of some related work concerning the previous proposal – discussed in the introduction. We have decided to maintain them due to their importance to the rationale that conducted us to the present proposal. They also contribute to give a broader perspective of the current debate concerning the relation of information and the topology of a social network.

2.1 Dynamics of Networks

Social systems are highly complex and dynamic. In this context, it is not trivial to understand how interactions among a large group of people result in a structured social group. When considering it as a social network, it is highly useful to represent it as a graph, wherein nodes stand for social entities and edges for their connections [73]. This approach has been used since the first sociogram displayed by Jacob Moreno [51] and this representation allows the usage of methods developed by the complex networks community [55, 6, 7, 9]. Complex networks come from the abstraction of complex systems as graphs, or networks. Complex systems encompass systems composed of components whose interactions lead to nontrivial behaviors, such that focusing only on the isolated components or a system's macro view does not provide enough understanding about it [5]. In the complex networks area, the connections between system's components are abstracted as the edges of a graph and this framework led to theoretical results, mostly by the physics community [9], which benefited research initiatives about social networks [18, 45, 30, 14, 74, 2]. Therefore, the works presented in this section are mostly results



(a) Time t_1 (b) Time t_2

Figure 2.1: Link Prediction Example

coming from the complex networks and social networks research.

One important topic in this context is to understand how social networks change over time. This is often addressed with online social networks data, which have interesting evolving behaviors – e.g., they usually are mostly growing [45] and can be highly dynamic [54], i.e., edges constantly being created and removed. Nevertheless, its evolutionary behavior cannot be explained only by random models as the one proposed by Erdös and Rényi [24]. Usually research in their temporal changes concern nodes and edges behavior and the simplification that relies only on the prediction of their measures have not obtained good results [50]. Before the advent of online social services, early works of sociology already had an interest in describing what would be the mechanisms behind the creations of connections in social networks [73, 32]. Usually, they exhibit a power-law degree distribution [9, 25], indicating that some individuals got much more connections than others. Barabási explained that the heavy-tailed degree distributions of several networks – usually denominated scale-free – would be the result of the *preferential attachment* effect [6].It postulates that the probability of a new edge arriving at a node is proportional to its degree. Sociologists have viewed this phenomenon as the Matthew Effect, or the "rich gets richer", as highly connected individuals have a higher probability of establishing new connections [61].

Despite the preferential attachment being capable of describing the existing pattern on the arrival of links in scale-free networks [13, 45], it is not enough to explain all the dynamics of links' creation in real social networks, as shown by Leskovec et al. [45]. They demonstrated that the likelihood of new links emergence is related to some network structures, e.g., it is usual for new links to close triangles in the network, connecting individuals that are two hops away. Furthermore, individuals behavior is also important to determine the appearance of new connections. They showed that nodes that have recently participated in a new edge, more probably will participate in new edges. The importance of this work is to show the presence of a feedback loop wherein the social network structure dictates its future structure.

The growth of complex networks is intrinsically related to the dynamics of their edges formation, as their structure is defined by their nodes and edges composition. The specific problem of predicting future edges in a network was formalized by Liben-Nowell and Kleinberg as the *link prediction problem* [46]. Figure 2.1 illustrates it: is it possible in time $t1$ to detect the creation of the edge between b and c at time t_2? Initially, the

prediction of links focused on which network structures could imply the emergence of new connections, i.e., if exists some arrangements of nodes and edges highly correlated with new edges creation in specific places. As mentioned before, one substructure that is important in this process is the triadic closure, i.e., links closing triangles. Furthermore, it is often the case that the likelihood of a link between two nodes is connected with the number of their common neighbors as it implies a high number of triangles being closed.

The link prediction problem is important for a high variety of topics, from finding missing links [20] to measuring nodes influence [62]. It has been explored in diverse types of settings, e.g., in collaborative networks [46], in multidimensional networks [67]. There are deterministic approaches [46, 67] to address the problem and it has also been treated as a learning problem [35]. Furthermore, works have shown that some characteristics of a pair of users increase the probability of them establishing a new connection – e.g., if they have similar interests [2], if they have high number of common friends [46], etc.. Current state-of-the-art approaches rely on random walks [41, 71] based algorithms that make use of nodes and edges information to accurately obtain the most probable links to be created [4, 34].

The link prediction is able to address the understanding of networks evolution at the level of individual edges. Another possibility is to look for subgraphs of given topologies which are correlated with the emergence of other subgraphs in the network evolution. The benefit of this more general view is that it allows comprehending the network evolution more thoroughly, considering appearance and removal of its nodes and edges. Tamm et al. [70] also explored networks evolution via their substructures, specifically, in terms of their motifs distribution. Motifs are subgraphs that exist in some networks with a higher frequency than in an equivalent random graph, i.e., a graph with some equal properties generated by a random model [9]. They conjectured that motif distributions in each network state could affect its evolution as the entropy of some states could leave them more stable than others. This study addresses the phenomena in abstract models and, to the best of our knowledge, this kind of test has not been verified in real networks.

The mining of important subgraphs to investigate networks evolution is an approach related to the techniques used in computer science. The mining of subgraphs in graphs have already been explored in generic approaches, e.g., gSpan [77] and Gaston [58]. Bringmann et al. [10] proposed an approach to mine subgraphs that are frequently correlated with the network structural changes from a collection of network snapshots . One example is a pair of nodes two hops away that will usually close a triangle in a future state. These frequent subgraphs can be considered patterns of evolution, expressed as *GER(Graph Evolution Rules)*. Figure 2.2 shows two examples of *GER* that the authors obtained in their experiments. The subgraph in *GER1* is highly correlated with the creation of an edge between the still not connected nodes two hops away, and the one in *GER2* with the appearance of a new node. These mined substructures could be used to predict the arrival of new edges and nodes in the network future graph.

Our initial proposal intended to understand the evolution of a social network taking into account the information that is shared by its users. To achieve this, we intended to use a subgraph pattern mining similar to the one proposed by Bringman et al. together with data of the motifs distributions. Unfortunately, this approach required longitudinal

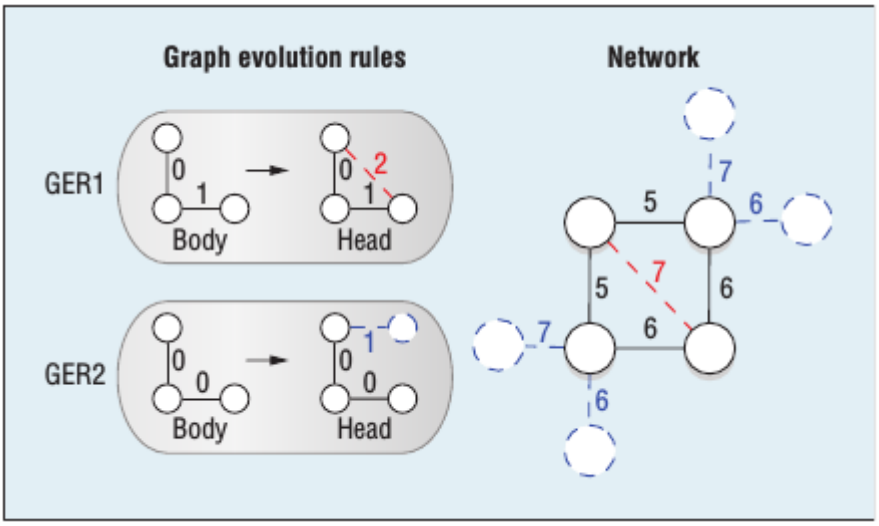


Figure 2.2: Graph Evolution Rules example. The weight of the edges denotes how long they exist. Extracted from [10].

data of nodes and edges, infeasible to be obtained. Our final work is concentrated in how the information that traverses the network is related to the users connections without concerning about endogenous changes in the network topology. The majority of works tries to understand connections mechanisms in terms of network topology, as the ones presented in this section. However, we are interested in how the information shared can also provide knowledge about connections. This has also been explored by other works. Thus, in the next section, we present works that explore the concept of information flow in social networks and how it is related with the network topology.

2.2 Information in Social Networks

The study of information diffusion is orthogonal to the temporal changes of networks structures. Human communication incorporates information, which may influence and shape people's behavior. Thus, sometimes it is considered analogous to the epidemic spreading [60, 12] phenomena. An individual that adopts a behavior or idea due to an received information is analogous to an infected individual. However, it is important to notice that the contagion by information focuses on the spread of social behavior, which has some specificities. For instance, the complex contagion [17] defines that people are more likely to acquire some behavior or join a cause if they are repeatedly exposed to whom already adopts it [16]. Romero et al. [66] explored how users adopt *hashtags* in *Twitter* and found evidence that social theories about spreading and adoption can be applied to online social networks. The effects of social contagion are easily seen in the role that online social services have taken in recent mass mobilizations, mainly in the recruitment process [30]. Based on those findings it is natural to assume that the spread of information may influence the social network topology, which, by itself, dictates how information flows on it. This feedback loop between topology and information is the main focus of this section as this relation is a core concern of this dissertation.

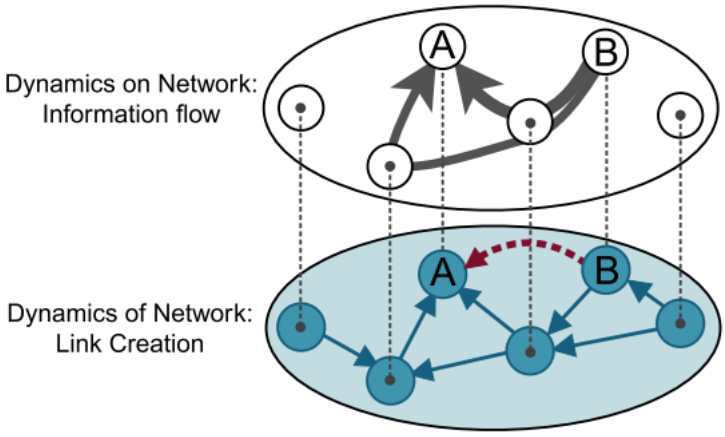


Figure 2.3: Dynamics of the social network structure and the flow of information. Extracted from [76].

There is no clear definition of information in a social network. In this work, we consider it as the different kinds of content that flow in a network and may affect people. Thus, information can be seen as a set of contents traversing individuals' connections, which can affect their opinions or ideas. This is analogous to the Bateson's general definition of information as composed of pieces that are supposed to be "a difference that makes a difference" [28]. In the Twitter social network, Myers et al. have shown that different kinds of information interact between themselves, cooperating or competing in the process of contaminating users [53]. Furthermore, they have shown that the semantic similarity of their content may indicate if they cooperate or compete with each other. Also considering Twitter, Weng and Menczer [74] proposed a method to model content as topics of information. They built a co-occurrence graph of hashtags and used a community detection method to find topics based on the assumption that semantically similar hashtags are likely to appear together in the same message. They found that active users attract others by beeing focused on a small number of topics, which makes them having more impact in the network with respect to the topics they are focused on. The importance of their approach to model information is that it is able to capture the latent relation among hashtags. This procedure is essential as hashtags do not have a hierarchical structure such as of tags or labels, which makes them not appropriate to be classified in taxonomies [36]. Other approaches to model information on Twitter required a manual step to annotate the data set [43] or needed an external ontology to do the semantic classification [44]. Therefore, this choice of using a community detection method to detect topics is also used in our work to model the information in our Twitter data.

Some recent efforts have been directed to the study of how the information flow influences the creation of links. Weng et al. [76] discern the dynamics *of the network* from the dynamics *on the network*. The former is the dynamics of networks structural changes, which can be considered the network evolution covered in Section 2.1. The latter is the information that is spread and traverses a network, i.e., the dynamics of the flow of information. Figure 2.3 shows one example, from the work of Weng et al., in which the flow of information affects the link creation process. A message created by user B that arrives at user A may influence the latter to follow the former. Their work was an extension of the work of Leskovec et al. [45] and with the addition of the information flow data they

described the rules that defined the link formation behavior. They presented evidence that the information flow has an important role in the process of link creation, around 12% of new edges were motivated by the information flow, indicating that the network evolution cannot be explained merely by its topological structure. Another important result, that is directly related to this dissertation, refers to the fact that, while some users create connections mostly based on friendship, others are more guided by the content that users produce and share.

The previously mentioned work of Gonzalez-Bailon et al. [30] explored the 15M Spanish social movement and found a rapid increase in the adoption of *hashtags* related to the social movement. One interesting question about this kind of phenomenon is if it is followed by a significant change in the connections structure. This issue was explored by Meyers et al. [54], who were interested in how the rise of abrupt changes in the information flow dynamics influence the creation and removal of links. Their work found that in a similar event, the "Occupy Wall Street" protest movement against income inequality, the cascade of tweets was likely to cause *follow* bursts, i.e., people start to follow others with the abrupt increase in the retweets of some contents. Actually, they could capture general bursts and they also verified that some tweets with offensive content caused the opposite, *unfollow* bursts. These *unfollow* and *follow* bursts generated a significant change in the network and often left the users' neighborhoods more similar to them, leading to a more cohesive network. This verification led them to create a model to predict which bursts of retweets created a new burst of *follow* connections in *Twitter* based on the intuition that users tend to connect to similar users.

The main contribution of these works is to give evidence that the information flowing in a network contributes to shaping its structure. Our initial proposal was tightly connected to their perspective and our intention was to explore the connections and information flow patterns to predict the future state of a social network via an approach similar to Bringman's et al. [11]. However, we changed the focus of the research due to difficulties in obtaining the required data. Instead of looking to the flow of information and network evolution, we analyzed if the cohesion of information is related to users connections, i.e., if connected users are generally more similar in the kinds of information they share. We address this issue through the perspective of topical homophily, a central concept of our research. Homophily is the subject of the next section.

2.3 Homophily

In a social system, individuals connect to each other driven by different mechanisms, from preferential attachment to creating shortcuts for the consumption of information[6, 76]. Sociologists have long believed that individuals are likely to establish relations with alike, what is known as the homophily tendency [42, 49, 38]. The classification and definition of homophily come in different flavors and often it manifests differently according to the traits considered in the analysis [49] – e.g., connected individuals may be highly similar with respect to religion and not so much with respect to sex. Furthermore, it is necessary to assess how much individuals are expected to be similar in the considered

environment, i.e., to assess the expected similarity among random assortments of individuals. This is crucial to determine if the similarity among connected individuals differs significantly from the expected similarity among random pairs of individuals. The difference between these two situations is captured by the two concepts of *baseline homophily* and *inbreeding homophily* introduced by McPherson & Smith-Lovin [49], a classification that is used in this research.

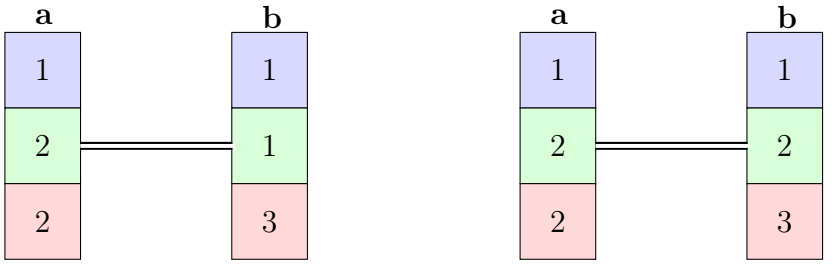
McPherson & Smith-Lovin defined baseline homophily as the expected homophily between random pairs of individuals from a population. It captures how the population is similar overall regarding some traits of interest. Inbreeding homophily stands for the deviation from the level of baseline homophily when considering the similarity of dyads, i.e., between pairs of connected individuals. By this definition, a highly homogeneous environment will have a distinct level of baseline homophily from a heterogeneous one, as in the former pairs of users are expected to be more similar than in the latter. Thus, it is impossible to assess inbreeding homophily without assessing the baseline beforehand.

Different traits can show different levels of the two kinds, for instance, in a given context, age homophily showed higher levels of baseline homophily and gender showed higher levels of inbreeding homophily [49]. The levels of inbreeding homophily may be a product of diverse factors. Individuals may have chosen to bond by their similarity or may have been induced by other factors. For instance, some characteristics of the social system may affect the opportunities of connection, restricting further individuals choice to bond with another. As the classification by McPherson & Smith-Lovin is not concerned with describing factors which generate inbreeding homophily, it is often used in empirical studies interested in homophily [65].

In settings where it can be verified, the classification of *choice* and *induced homophily* [48, 65, 38] can nicely demonstrate the underlying mechanisms of connections. Choice homophily is attributed to dyads wherein the similarity is a factor that determined the individual choice to establish the connection, i.e., there is a causal relation between the similarities and the individuals' preference to bond. Induced homophily categorizes the situations in which the similarity of the dyad is a by-product from the opportunities of connection. Thus, the similarity between connected individuals is not a factor of choice but induced by the social structures. For instance, suppose that in a setting boys are be induced to make friends with other boys, thus their friendship will not be a matter of choice. The dichotomy presented in this classification is interesting because it can quantify how much homophilous connections are a result of individual's psychological preferences.

Another important concept that can be considered a part of inbreeding homophily is the believed tendency that connected individuals become more similar between themselves over time, also known as *social influence* [3, 23]. Social influence theoretically defines that, in a dyad, there is a transmission of individuals attributes that are possible to change in the time scale of the study, for instance, political opinions or religion. As in our work we are not focused on disentangling these three processes – namely, social influence, choice and induced homophily – we preferred not to consider them in our analysis.

Besides the importance of studies that look for evidence demonstrating how homophily is manifested in real scenarios [48, 49, 38, 40], homophily is also an important premise for theoretical social models. The model proposed by Robert Axelrod to describe the



(a) Vectors before interaction. (b) Vectors after interaction.

Figure 2.4: Illustration of the process of interaction and social influence in the Axelrod model. Vectors interact because they have, at least, one characteristic in common. After the interaction, they become more similar to each other.

dissemination of culture is probably the most important in this respect and it is basically defined over two premises: choice homophily, *"similarity begets friendship"* and social influence, *"friendship begets similarity"* [3, 15, 14]. In the Axelrod model, culture stands for the set of attributes that characterize an individual and which are subjected to social influence – e.g., language and religion. The model abstracts the attributes as a set of discrete variables represented by a feature vector F, wherein each position stands for a different attribute. Moreover, each attribute can assume a value according to the global parameter q, which defines the number of possible traits that each characteristic can have. Thus, the number of features $|F|$ and the number of traits q establishes how the culture is discretized.

In the Axelrod model, individuals interaction is defined by choice homophily, specifically, they interact with probability proportional to the number of features wherein they have the same trait. When this interaction takes place, one attribute in which the two differ becomes the same. Strictly speaking, the trait of one's feature is passed to the other, leaving the two individuals more similar as a result of social influence. The vectors and this process are illustrated in Figure 2.4.

This process may induce one to think that the result would be a homogeneous system, i.e., a system in which all individuals have the same culture vector. However, the number of homogeneous regions – regions wherein all individuals have the same culture – vary considerably according to the parameters $|F|$ and q, indeed, the number of cultural regions increases according to the number of traits [14]. We adopted an analogous approach to model individuals in our study, however, our approach is not concerned with modeling individuals according to their whole culture, we model them in respect to their affiliation in different topics of information.

Our dissertation explores topical homophily in online social networks. Topical homophily, here, is addressed using topics of information found by the same method of Weng and Menczer [74]. A similar work was conducted by Aiello et al. [2] in the context of tagging social networks (Flickr, Last.fm, and aNobii) using the tags attached to items. In these networks, the tags are used to classify resources, a different usage than hashtags on Twitter. In their approach tags were used directly to assess users similarity and they found that users topical similarity are related to their shortest path distance on the social graph. Moreover, the measured similarity allowed them to predict some links of the

graph. Crandall [23] explored the extent to which the process of selecting most similar users to establish relations leads to an even higher similarity by social influence. They used Wikipedia and LiveJournal datasets – article and blogging based networks – and modeled users in vectors similar to the ones proposed by Axelrod [3] using their history of editions. This allowed them to verify that, after users interact through the selection process the similarity among them tends to increase, which provides evidence of the social influence principle. Ciotti [19] looked for homophily in citation networks. In their setting, they analyzed articles similarity considering their bibliographies items as attributes and their main contribution is to provide a way to assess similarities among articles and recommend missing citations.

These works are more related to networks centered in some kind of digital artifact, e.g., image, article, etc. In our work, we are not interested in artifacts, but in the relationships of individuals. Furthermore, we believe that hashtags or other features, by themselves, are not sufficient to assess the similarity among users as they do not capture the latent semantics present on the sharing of information. We believe that it is necessary a higher granularity to really capture users affiliation in different kinds of content. Thus, we chose to model the topics of information to assess users homophily. We are mainly interested in understanding of whether it is possible to verify levels of baseline and inbreeding homophily in Twitter connections and how much users relationships are affected by the topics that they are affiliated with. Our experiments and results are described in the next chapters.

Chapter 3

Twitter Data

We have chosen to use Twitter as our online social network data source due to its information-driven nature. Furthermore, its data is relatively feasible to be obtained. In this chapter, we give a brief description of Twitter and the process of data gathering and cleaning. In the subsequent chapters, we demonstrate how the data analysis was conducted.

3.1 The Twitter Social Network

There is no clear definition of social media or online social network, however, there is a general consensus that services like Twitter[1] and Facebook[2] are instances of social media services [59]. One of the main characteristics of an online social network is that their users are represented by profiles, which allow them to connect with other users or groups of users. These connections have different meanings across the different kinds of social medias. As our research is concerned with Twitter, we describe its types of connections that might exist between users that are relevant to this project:

Follow: when a user decides to follow another, the former will receive all the tweets that are shared by the latter in her 'feed'. If a user a follows a user b, by Twitter definition, the user a is said to be a *follower* of the user b, and the user b is said to be a *friend* of the user a. This definition is important as the connection in Twitter might not be reciprocal. Figure 3.1 illustrates the connections as black edges. The orientation of the edges is according to the flow of information, i.e., messages are created by the friend and arrive at the follower, so the edge goes from the friend to the follower.

Mention: a user mentions another through the convention *'@ + username of the other user'*. Mentions might denote a bigger interaction between users than *follow* relations They might emerge from conversations (*Replies*), they can be an endorsement and propagation of ideas (*Retweets*), and a mention is often used to grab users attention. Figure 3.1 illustrates them by the red dashed edges, edges leave the node responsible for the mention and arrives at the mentioned node.

[1] http://www.twitter.com, accessed in September 2016.
[2] http://www.facebook.com, accessed in September 2016.

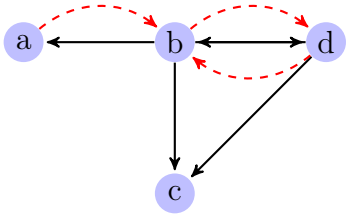


Figure 3.1: *Follow* and *mentions* connections. Follows and mentions are represented by the black and red edges respectively.

3.1.1 Information on Twitter

Twitter is an information-driven social network and information flows in it through short messages of 140 characters called *tweets*. This specificity is important as the nature of their messages are shaped by this limitation of size. They have to be concise and they often use tokens that refer to a specific subject. Undoubtedly, *hashtags* are the most important type of token used by users on Twitter. They can categorize messages and indicate to other users the subjects of the message or how the user is positioned about the subject, e.g., a hashtag may make explicit the sarcasm of the message. The importance of hashtags lies in their ability of make explicit the topics to which the message belongs. Figure 3.2 shows a tweet example.

So nice having the sis over at the flat for the night :) #munchies #movies #spooning

Figure 3.2: Tweet containing hashtags. Extracted from our dataset.

The use of hashtags is interesting for researchers and data scientists as they may be used as a guidance in the data collection and selection procedures while analyzing a group of users in Twitter [31]. Specifically, the most common way of obtaining Twitter data is by its public API[3] and hashtags are often used to filter messages that belong to the topics of interest [31]. In this work, we collected tweets without the hashtags filter, nonetheless, we used hashtags to build topics of information as we will describe in Chapter 4.

3.2 Data Gathering and Data Filtering

This research was conducted jointly with the COSNET Spanish laboratory and our first dataset was obtained by them in two steps. A second complementary dataset was further collected based on the existing data. The data were obtained through the public Streaming[4] and REST APIs[5] of Twitter. The process was centered around tweets from the United Kingdom and Ireland. We describe each step in the next paragraphs and a summary of the data is shown in Table 3.1.

Tweets Obtainment Using a geolocation filter, tweets from the United Kingdom and Ireland were obtained trough the Twitter Streaming API. From all the users in the dataset,

[3]https://dev.twitter.com/rest/public, accessed in September 2016.
[4]https://dev.twitter.com/streaming/overview, accessed in September 2016.
[5]https://dev.twitter.com/rest/public, accessed in September 2016.

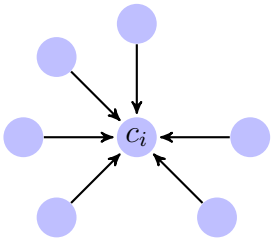


Figure 3.3: Illustration of the star structure of the central users.

Data	Raw
Tweets	98,506,315
Tweets with Hashtags	16,935,625
Distinct Hashtags	4,320,429
Users with Tweets	1,286,816
Users with Hashtags	774,596
Central Users	9,632
Central Users' Friends	4,190,244
Central Users in Second Crawling	6,296

Table 3.1: Crawled Data Summary

10,000 users with more than 100 tweets were selected. This was an empirical decision to select the users that were relatively active in the social network. Furthermore, the set of tweets of these users were complemented using the Twitter Search API (a component from the REST API). The final dataset has 98 million tweets from January 18th to September 2nd, 2013. Of these tweets, almost 17 million contains some hashtag and there is a total of 4 million distinct hashtags.

First Crawling of Friends We denote the set of $10,000$ users selected in the last step as the **central** users. Our analysis is centered around this set of users and their friends – i.e., the users they followed – were crawled through the REST API. This resulted in a set of star networks as illustrated by Figure 3.3. The user in the center, c_i, is a central user and the others are users that she follows. The edges follow the direction of the flow of information. The process of crawling the friends ended on June 26th of 2013 and resulted in a set of 4 million followed users.

Second Crawling of Friends One limitation of the original data was that it did not have information of the reciprocity of the connections. Furthermore, it is interesting to have information about the edges that have endured for a significant amount of time. Hence, on February 23rd of 2016, we started a new crawling of the central users' friends in a similar fashion of the previous step and, in addition, we crawled also the users who were followed by each central user friend. This lead to a total of 141 million friends relationships and the set of central users were reduced to $6,296$ users, as some of them were not available anymore. These set of connections were only used in the analyses described in Sections 5.2.3 and 5.2.5.[6].

[6]More information about the crawled follow connections are shown in Appendix B.1

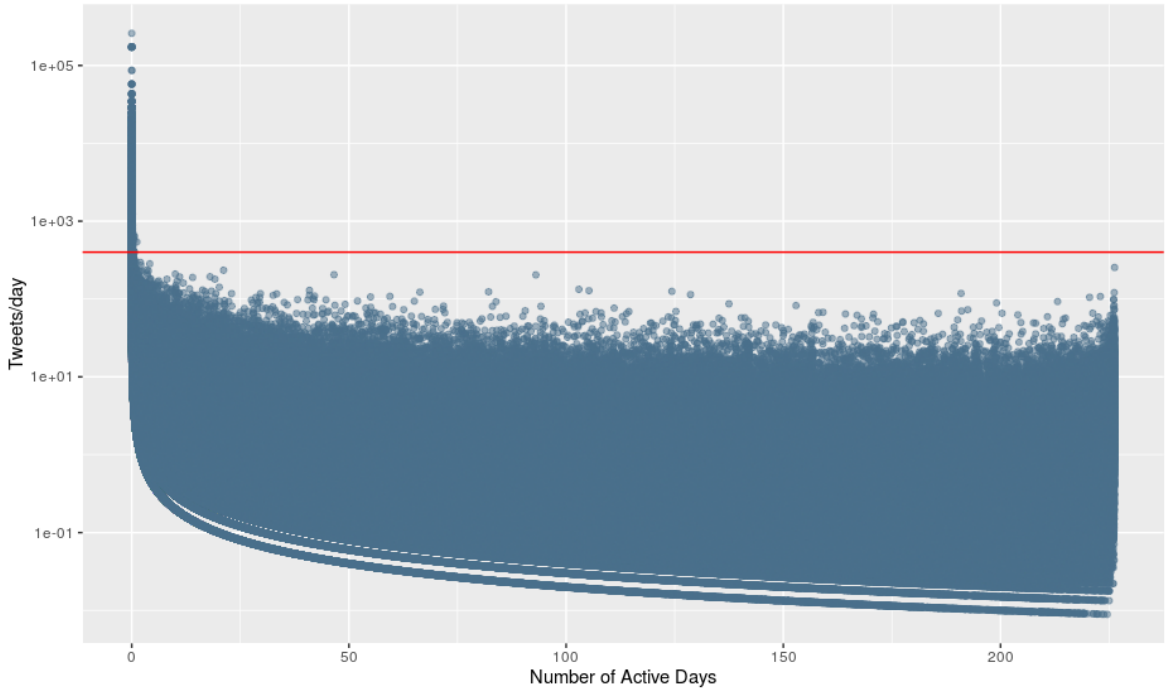


Figure 3.4: Users activity according to their tweets.

3.2.1 Selection of Users

Our work is centered around topics of information, which are composed of hashtags and we describe the process of their construction in the next chapter. The users considered in our analysis had, at least, one tweet with a hashtag in order to assess which topics of information they were affiliated with. Thus, we selected the $774,596$ users from the 1 million of users with tweets.

Before starting the analysis, we removed the users that have been active for less than one day and the ones that might not be a real person, i.e., users that might be a bot. Here we define the users active days as the number of days between their first and last tweet.

To uncover which users might be bots, we first looked to their number of tweets per day and their number of active days. Considering only the users that were active for, at least, one day, Figure 3.4 shows a scatter plot of the users' number of active days versus their ratio of tweets per day. This figure shows that there is a significant number of users that have been active for only a few days, having a large number of tweets per day. Besides them, most of the users are uniformly distributed. We decided to remove the users that had more than 400 tweets per day, as we consider that is normally infeasible for a real person to produce this quantity of tweets. This empirical threshold is shown by the red line in Figure 3.4 and this process resulted in a set of $693,953$ users.

The data produced in this stage is the basis for our analysis and sufficient for the detection of the topics of information. After the detection process, the number of users considered in our analysis was further reduced, because the considered hashtags were reduced. The next chapter details how the whole process was conducted.

Chapter 4

Topics of Information

This dissertation addresses how the users' communication of different kinds of information is related to their connections. In order to represent these different kinds of information, we modeled them as topics based on detected communities of hashtags. Section 4.1 describes how the model works and Section 4.2 details the process of topics detection from our Twitter data.

4.1 Model

Information in Twitter is traversed through tweets. They are short messages with a highly dynamic vocabulary, which makes traditional text clustering techniques not suitable. Thus, we decided to exploit the hashtags social annotations, described in Section 3.1.1, which are present in Twitter messages. They are used by us as a guidance in the process of collecting topics of information, which is the approach used by us to model the information generated by Twitter communication. The formulation of topics of information enables them to capture the latent semantics of the messages through hashtags co-occurrence. Topics are defined via a community detection method, which we introduce in Section 4.1.1 and subsequently we describe the model behind the topics of information in Section 4.1.2.

4.1.1 Community Detection

Community is the most studied structure of networks because it can capture subnetworks that present distinct properties and configurations [57, 26] and nodes inside a community are believed to share similar properties or have similar roles [26]. Communities are dense subgraphs in relation with the whole network, i.e., nodes inside a community have a higher proportion of edges between them than with nodes outside the community. Community detection is analogous to the to graph partitioning problem in computer science [37], an NP-complete problem. Some of the most used methods are based on hierarchical clustering [73] and modularity optimization [56].

We use a community detection method in our project to find coherent communities of hashtags that we categorized as topics. More details are in Section 4.1.2. We adopted

the OSLOM[1] tool [39]. OSLOM works by the perspective that groups of densely connected nodes that are just a product of random fluctuations should not be considered as communities. To verify if this is the case, it has a fitness function to evaluate the communities by comparing the probability of finding them in a random null model. If a cluster is highly probable to occur in a random configuration, it should not be considered a relevant community. It finds clusters by a method which starts from random nodes and adds the nodes that will build the most relevant clusters. This evaluation can be also used to verify communities detected by other algorithms and the OSLOM tool allows the execution of the Infomap [68], Louvain [8], and Copra [33] methods as input, returning the best detection found. Furthermore, OSLOM is a local optimization method and does not suffer from the problems of global modularity optimization [27]. It also has the advantage of beeing highly flexible, being possible to be executed in weighted and large graphs, and can detect overlapping communities, which is desirable in our case.

4.1.2 Co-occurrence Graph of Hashtags and Topic Detection

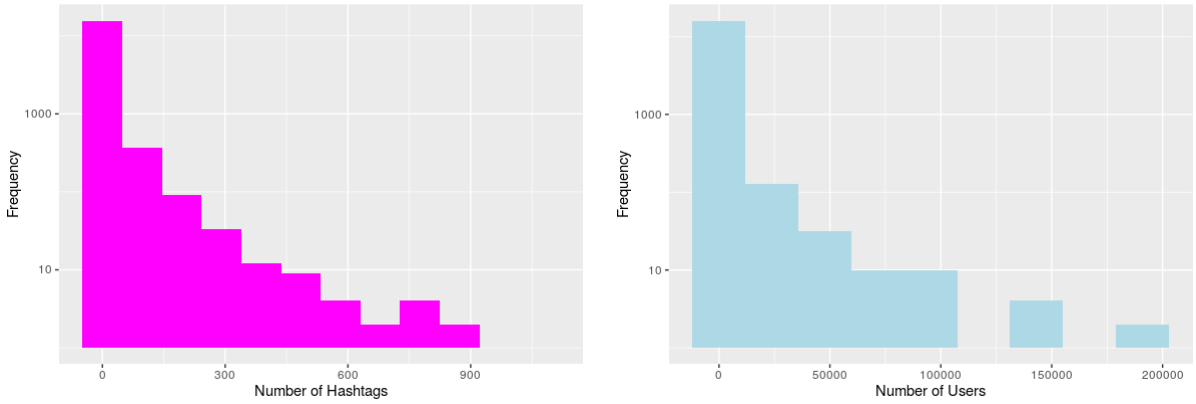
In our analysis, we built topics of information considering tweets with hashtags, as they are indicators of the tweet content. Furthermore, it is common for users to insert more than one hashtag in a tweet, and we exploit this aspect to build a semantic mapping of Twitter messages. We assume the existence of a semantic association between hashtags that co-occur in tweets. This is analogous to the assumption that words are semantically associated if they are likely to co-occur frequently [72]. The use of this assumption makes our method focused only on the implicit semantics given by Twitter messages, i.e., it does consider explicit semantics given by other sources . This semantic mapping is captured by a weighted co-occurrence graph of hashtags, which we built by extracting all pair of hashtags that co-occurred in each tweet in our dataset. In this graph an edge (h_i, h_j) describes that the hashtags h_i and h_j co-occurred and, as the graph is weighted, $w(h_i, h_j)$ gives the number of different tweets in which they co-occurred.

The importance of the association of hashtags in the co-occurrence graph is that it allows the extraction of higher level semantic structures. We consider that the topics of information are sets of hashtags clustered together in the graph. Thus, we expect that they will reflect the higher level structures that emerge from the latent semantic association of hashtags. It is natural to see that these clusters could be captured by a community detection method and we decided to use the OSLOM tool [39]. OSLOM is able to capture for overlapping communities, a desirable feature considering that one hashtag may be used in different contexts.

We also tried our execution with the Link Communities method [1], however, as our network is very large and this method is more computationally intensive, we preferred to use the OSLOM tool. Moreover, previous work has verified that different choices of community detection method did not significantly impact the topical clusters found [74, 75].

This approach of building a co-occurrence graph and using a community detection method to find topics was also used by Weng and Menczer [74] through the Louvain

[1]Available at http://www.oslom.org/



(a) (b)

Figure 4.1: Semi logarithmic histograms of the number of hashtags and number of users in the topics.

method [8]. They assumed, based on the topical locality assumption, that semantically similar hashtags would appear in tweets together. Despite this being close to our premises, we do not presume that hashtags are similar, only semantically associated. Besides this remark, our proposal to find topics of information is roughly equivalent to theirs.

4.2 Topics by Community Detection

This section presents how we modeled the topics of information. First of all, we built a hashtag weighted co-occurrence graph to find communities as described in Section 4.1.2 using the 16, 935, 625 tweets with hashtags belonging to our dataset. As we removed hashtags that did not co-occur with any other, the co-occurrence graph resulted in 2, 090, 971 from the total of 4, 320, 429 distinct hashtags. In this graph, the edges represent a semantic association between hashtags, however, hashtags might have co-occurred in a tweet only by chance and without having a significant association. To reduce this noise, we removed all the edges between pairs of hashtags that co-occurred in less than 3 tweets, i.e., we removed the edges with a weight smaller than 3. This process led our final co-occurrence graph with 104, 308 hashtags. As mentioned in the previous chapter, users had to have at least one hashtag in order to be analyzed, thus, this reduction in the number of hashtags led to a final set of 608, 899 users from the previous total of 693, 953.

With the final co-occurrence graph of hashtags, we were able to find the topics of information with the OSLOM tool[2]. The application of OSLOM resulted on 2, 074 communities and 14, 118 homeless nodes, hashtags that did not belong to any community. We considered the communities and the homeless nodes as topics. Despite the latter possibly not significantly benefiting our future procedures, we believe that a hashtag alone can also carry information. Furthermore, increasing the topics should not affect the way we assess topical similarity among users, as it is later shown in Section 5.1.2. Thus, we consider a total of 16, 192 topics in our analysis. Figure 4.1a shows the histogram of the number of hashtags in each community with a logarithmic scale on the y-axis. There is a peak near 0 that corresponds to the 14, 118 hashtags that were not assigned to any community. The

[2]See B.2 for more information about our execution parameters.

other topics composed of the 2,074 communities had an average of 46 hashtags. Figure 4.1b shows an analogous histogram for the number of users in the topics, i.e., the number of users that had tweeted, at least, one hashtag belonging to each topic. Most of the topics have a small number of users, probably because most of them are composed of only one hashtag. The mean number of users in the topics is of 622.

Figure 4.2: Word clouds with the hashtags of 8 communities. Hashtags in each figure have their size proportional to their edge degree in the subgraph corresponding to the community.

Although there is not an easy way to ground the accuracy of this approach, we believe that it is a sound method for assessing the topics of information. Its premises and procedures are well defined over the semantic associations of hashtags. Furthermore, as is illustrated in Figure 4.2, the content of the topics appear to have a semantic sense. It shows hashtags clouds of eight different topics and they show the existence of consistent semantic relations in the topics. Further analysis could better verify the precision of our

approach.

In the next chapter, we describe how these topics are used to compute users topical similarity, which enabled us to analyze users' topical homophily in Twitter.

Chapter 5

Topical Homophily

Our work seeks to understand homophily regarding the topics of information that flow in a social network. We do this by modeling users according to their affiliation to topics and computing their similarity. This chapter verifies the existence of topical homophily and how it can provide information about relationships in the network. It details the process and shows the experiments and their results using a Twitter dataset. The results indicate that ties between users are likely to show higher topical similarity, which tends to increase with the ties strength. Section 5.1 describes how users are modeled through topics and the method to compute their similarity; Section 5.2 shows the results of the tests of topical homophily between users; Section 5.3 explores if a higher similarity is indicative of a higher number of mentions between connected users; and Section 5.4 explores if the similarity between users can indicate their connections.

5.1 Users Modeling and Similarity

5.1.1 Users Representation

In our work, individuals are Twitter users and we explore their homophily through their topical affiliation; what we will further refer to as topical homophily. In our analysis, we are not concerned with sociodemographic characteristics. Besides they being often not available or having a dubious veracity, we are interested in assessing users similarity with respect to topics of information. A user, in our model, is represented by a feature vector $\mathbf{u}$, which comprises her affiliation to all topics of information. The process of building a user vector is illustrated in Figure 5.1. The feature u_i corresponds to her affiliation in the topic i and has its value according to the number of hashtags belonging to t_i(the set of hashtags belonging to the topic i) that were used by the user in her tweets. As the communities obtained by OSLOM may overlap, the same hashtag may be computed in more than one feature. In this case, each hashtag adds a proportional value to each feature it belongs . The value of a feature u_i is given by Expression 5.1.

$$u_i \leftarrow \sum_{\{h \in H : h \in t_i\}} \frac{m_U(h)}{|\{t \in T : h \in t\}|} \tag{5.1}$$

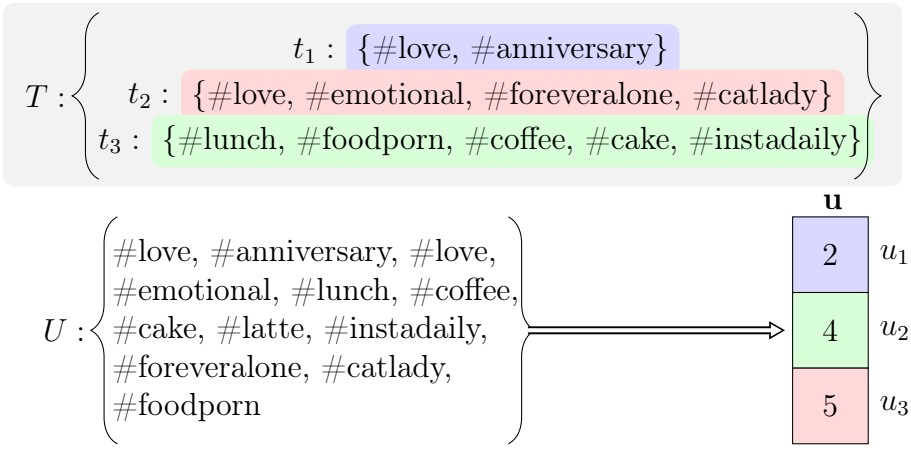


Figure 5.1: Building of a user feature vector.

All the hashtags used by a user are contained in a multiset $U = (H, m_U)$, wherein H is the set of used hashtags and m_U gives the number of occurrences of each one. T is the set of topics, i.e., communities of hashtags. Strictly speaking, each element $t \in T$ stands for a topic and it is a set containing the hashtags inside one cluster built by the community detection method. Figure 5.1 illustrates a user multiset and its transformation in the user feature vector via Equation 5.1. As *#love* appears in the topics t_1 and t_2, it adds 1 to their respective features.

5.1.2 Computing Similarity between Users

With the representation of users as feature vectors, we are able to compute the topical similarity between two users using as metric the cosine of their vectors [72]. The cosine similarity fits well to this task as it only focuses on the angle between vectors – i.e., it does not consider their length. The cosine similarity ranges from 0 to 1; identical users would have similarity 1; users that do not share anything in common 0. It is given by Equation 5.2. In our preliminary analyses, we also tested Kendall's tau, Spearman's rho and Jaccard similarity measures. We did not adopt them as they did not present significant difference and improvement with respect to the cosine similarity.

$$sim_{cos}(u, v) = \frac{u \cdot v}{||u||||v||} \tag{5.2}$$

Our definition of users feature vector considers that all topics have the same weight, i.e., the values of the respective features is directly derived from the number of hashtags used. This may be not suitable for our task as some communities of popular or general use hashtags should have a smaller weight. To overcome this, we establish that features shared by a large percentage of the users ought to have a small weight, likewise, features possessed by only a small percentage of users ought to have a large weight. The intuition behind this is that features corresponding to rare topics should be more discriminative of the topical proximity of users than features corresponding to frequent topics. Strictly speaking, we would like to take into account the information content of the each community [22]. Expression 5.3 describes the weighting scheme, which is defined in a TF-IDF fashion [72].

$$u_i \leftarrow u_i \times log\frac{|I|}{|\{v \in I : v_i > 0\}|} \tag{5.3}$$

I is the set of all individuals, i.e., Twitter users. For each feature i in a user vector, this method will weight the value according to the number of users that also used it – e.g., a feature that is shared by all users will have its value set to 0, which is desirable because it probably does not provide information for discerning users

5.2 Topical Homophily of Users

One of the first hypothesis verified in this work was if there was a higher similarity between the friends of a user than between randomly chosen users. As detailed in the previous section, similarity means topical similarity and users are represented by a feature vector of topics wherein each position stands for a topic. In Section 4.2, we built a total of 16,192 topics using the hashtags extracted from the tweets and, in this process, the number of hashtags was reduced to 104,308. Therefore, some users had to be removed as they have to have at least one hashtag belonging to a topic in order to be represented by the vector of topics. This also reduced the number of central users and also the number of users that were followed by each central user. We further describe three sets of users as they are the final set of users used in the subsequent analyses:

Population: a set with 608,899 users that had tweeted, at least, one hashtag belonging to a topic;

Followed: a set with 214,089 users that had tweeted, at least, one hashtag belonging to a topic and were also a friend of one of the chosen central users;

Centrals: a set with 9,490 users that had tweeted, at least, one hashtag belonging to a topic and were also a central user, details are shown in Table 3.1.

Naturally, the set *Population* contains the sets *Followed* and *Centrals* and its users are the population considered in our analysis. Each individual was represented by a vector wherein each position stands for a topic found by the method described in Section 4.2. To reduce the value of topics that were shared by a lot of users, we weighted the users' vectors by the process mentioned in Section 5.1.2. In this process, we weighted the topics according to the number of users in the set *Population* that were affiliated with each of them.

5.2.1 Assessing Homophily

Our hypothesis that users are less similar to random users than to the ones they are connected will be addressed here in terms of the *baseline homophily* and *inbreeding homophily* classification introduced by McPherson & Smith-Lovin [49]. Here, we considered baseline homophily as the expected average similarity between users and others from a

random group of the population. Inbreeding homophily here is captured by the difference between the baseline distribution and the distribution of averages of similarities between the users and those with whom they form a dyad. A dyad, here, may be formed by a *follow* or *mention* relationship. The definition of baseline and inbreeding homophily are given by Definition 1 and 2, respectively.

Definition 1 ***Baseline Homophily*** *is the distribution of averages of similarities of considered individuals with random users of the population. Each value of this distribution is constructed as follows:*

For each considered user, a random group of users from the population is selected. This random group is of the same size as the number of relationships the considered user has. Then the similarity between the considered user and each one of the random group is calculated. Finally, the average of the similarities is computed.

Definition 2 ***Inbreeding Homophily*** *is the deviation from the baseline homophily when considering the similarity of the dyads. Thus, to assess the inbreeding homophily is necessary to build the distribution of averages of the dyads similarities. Each value of this distribution is constructed as follows:*

For each considered user, the group of users that are in a dyad with her is selected. Then the similarity between the considered user and each one of this group is calculated. Finally, the average of the similarities is computed.

The deviation is captured by two tests. First, we assess the degree to which the distributions differ by the Kolmogorov-Smirnov[1] *test. Then, the likelihood of the distribution of dyads yielding higher (or lower) values of average similarity is captured the Mann-Whitney U test*[2]*. For both tests, a* $p - value$ *is also calculated to assure statistical significance.*

5.2.2 Homophily on Follow Relationships

In this section, we initially explore the inbreeding homophily with respect to the *follow* connections. Our hypothesis is that users are, on average, more similar with their friends, i.e., we expect the inbreeding homophily to be significant. This hypothesis is explored through Definitions 1 and 2. Strictly speaking, our hypothesis is that the distribution of similarity averages of the individuals with their friends would be yield higher values than the distribution of averages with randomly chosen individuals from the population. We tested this hypothesis using the central users and their friends, we show the results in Figure 5.2.

Figure 5.2 shows three histograms[3]: *Friends*, a distribution of averages computed for each central user with her friends; *Followed* and *Population*, distributions wherein, for each central user, averages have been computed with a group composed of randomly chosen users from *Followed* and *Population*, respectively, and this group has the same size of the set of central user friends. We also used the set *Followed* because users in this set are guaranteed to be followed by someone, implying in a difference from the whole

[1] See A.2.

[2] For a brief explanation of the Mann-Whitney U test and the common language effect size, see A.3.

[3] The histograms with *'Density'* in the y-axis are shown in density scale, see A.1.

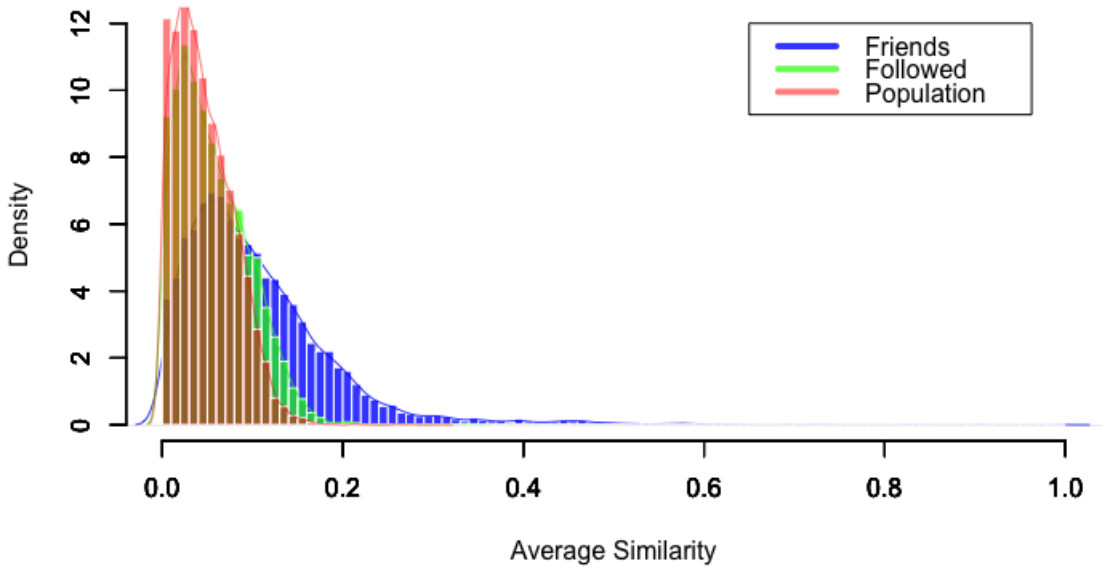


Figure 5.2: Distributions of the averages of similarities between the central users and their friends and between central users and random groups from *Followed* and *Population*

.

population. However, as will be shown, there is no significant difference. Therefore we did not use this set in the subsequent analyses.

As can be seen, all the distributions are concentrated in low values of the cosine similarity spectrum , i.e. $[0, 1]$. We consider that this effect is a result of the large quantity of topics and does not impact our results.

There is an overlap among the distributions, mostly concentrated in lower similarities. However, it is clear that there is a difference between both random distributions and the friends distribution. Strictly speaking, the Kolmogorov-Smirnov statistic between the distributions of averages with friends and with randomly selected users from *Followed* and from *Population* is 0.27 and 0.37, respectively, and both have a p-value $< 2.2^{-16}$. We also used the Mann-Whitney U test to verify if the distribution with friends were likely to have a higher average similarity than the two others. The results were positive with an effect size of 0.7 when comparing with the distribution of *Followed* and 0.75 with the distribution of the whole population, both with p-values $< 2.2^{-16}$. Furthermore, the medians of the distributions with *Friends*, *Followed* and *Population* were 0.087, 0.05 and 0.041, respectively.

This analysis shows that, on average, users tend to be connected to whom they are more similar. Strictly speaking, the similarity between friends is higher than the baseline similarity, what shows the presence of inbreeding homophily. This implies that a user tends to have a stronger topical similarity with friends than with randomly chosen users.

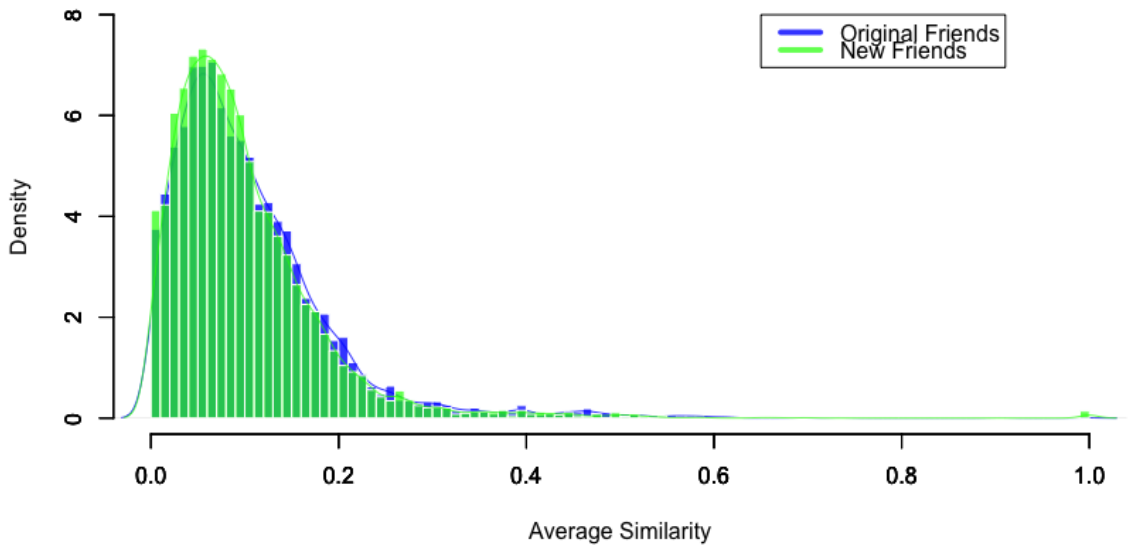


Figure 5.3: The distributions of similarity averages of central users with their friends from the two crawlings

5.2.3 New Connections

As mentioned in Section 3.2, we performed a second crawling of central users friends on 2016. We will further refer to this obtained *follow* relations as the new connections. Furthermore, we also collected the users that were followed by each central user's friend to explore the reciprocity of connections, which is covered in Section 5.2.5. One particularity of this dataset, as the connections of Twitter are highly dynamic, is that there might be a significant change in the connections pattern. We addressed this issue looking if the distribution of average similarity with the new friends in the new connections differed significantly from the distribution with the original friends, the result is shown in Figure 5.3. The set of new friends comprises all central users friends present in the new connections and which belonged to the set *Population*. As said in 3.2.1, the set of central users was reduced to $6,296$ users as the others could not be obtained in the last crawling. The two distributions almost totally overlap and the Kolmogorov-Smirnov test statistic between them is of 0.046 with a p-value of 2.209^{-6}, which indicates that they are roughly the same. Thus, we conclude that, regarding the pattern of topical similarity, there is no significant change between the original and the new connections.

Although the distributions are very alike, connections might have changed significantly during the interval between the two crawlings. Furthermore, if the topical similarity between friends is related to their connection strength, we expect the persistence of connections to be influenced by their topical similarity. We framed a hypothesis that the distribution of similarity averages of the connections that lingered have higher values than the connections that have not lingered. Thus, we expect that the central users' connections that were maintained had a higher average topical similarity in the moment of the first crawling then the ones which were not. We show the test of this hypothesis

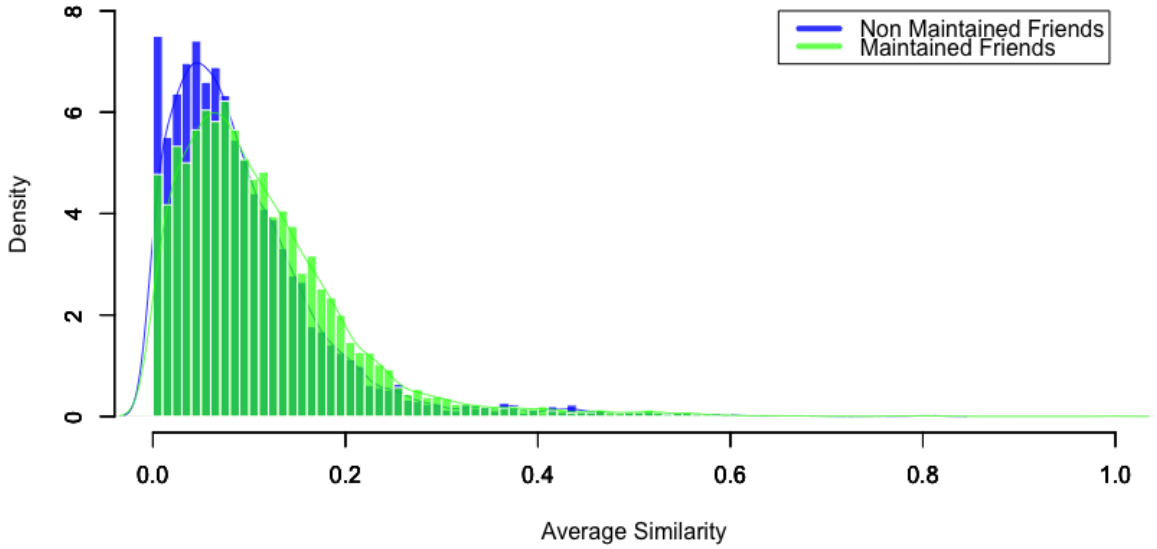


Figure 5.4: Distributions of averages of similarities betwen central users and their friends who were still followed in 2016 and between the friends that were not followed anymore in 2016. The central users considered here are the ones that had friends in the both situations, a total of 6,157 users.

in Figure 5.4. It shows the distribution of similarity with the friends that were still followed in 2016 and the distribution with the ones that were not followed anymore. The Kolmogorov-Smirnov test between the two distributions yields a statistic of 0.11 with a p-value $< 2.6^{-05}$, it is visible that the two distributions have a difference, however, it is difficult to conclude that it provides enough evidence to support our hypothesis that persistent connections have a higher similarity.

5.2.4 Users Interactions

Users on Twitter can use the convention *@username* to mention another user in a tweet. The interactions that happen through mentions are often seen as a relationship stronger than the *follow* connections [66, 31]. One hypothesis that emerges from such affirmation is that the topical similarity between mentioned users tends to be higher than between followed users. To test this hypothesis, we verified if the distribution of similarity averages with the mentioned users tended to be concentrated in higher values of similarity than the distribution of similarity averages with friends. Figure 5.5 shows the distributions obtained by the similarity averages between central users and the user mentioned by them and between central users and their friends. The distributions are roughly the same, the Kolmogorov-Smirnov statistic between them is 0.06 and the p-value is 2.2^{-15}. This does not bring enough evidence to contradict the general belief that there are stronger bonds between users that interact, since the comparison is limited to a specific kind of similarity (topical similarity). However, one is not able to say that the *mention* relations have a significantly higher inbreeding topical homophily than the connections with followed users.

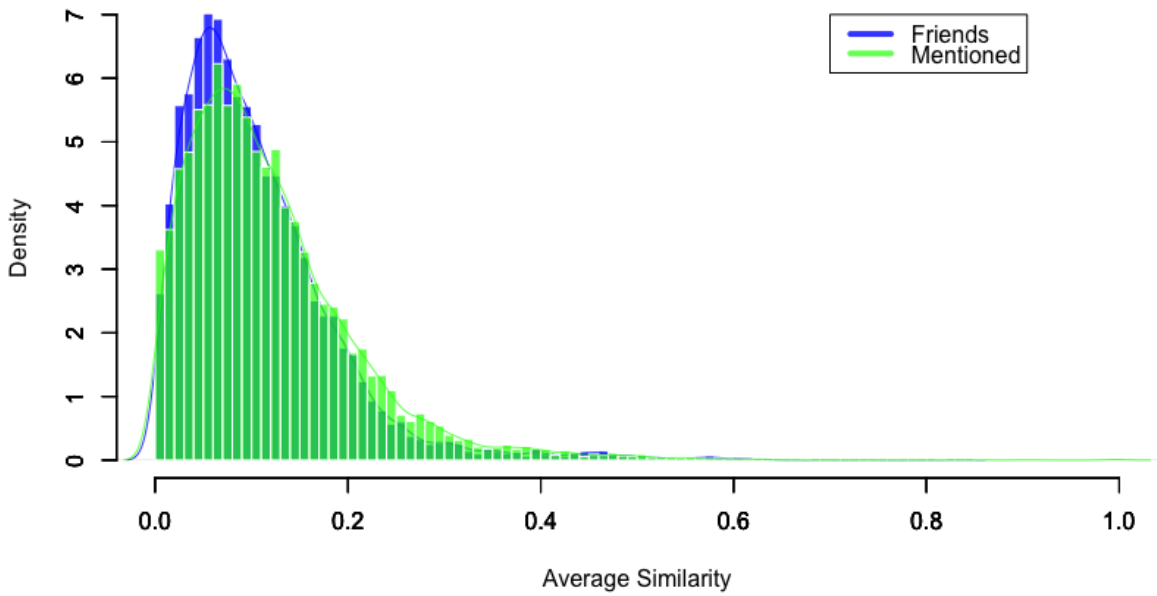


Figure 5.5: Distributions of similarity averages between the users followed and between users mentioned by the central users.

Both *mentions* and *friends* histograms show that most of the averages fall into low values of similarity and there is a positive skewness – i.e., skewed to the right – of the two distributions, what is not evident in the distributions with random users in Figure 5.2. Given the proximity between the two distributions of Figure 5.5, one question that emerges is if users, on average, follow and mention others in a close similarity pattern, i.e., if the users average similarity with friends and with mentioned users are correlated. We verified this correlation in Figure 5.6. The Pearson correlation between the two variables is 0.84, indicating that users that tend to follow similar users, also tend to mention similar users.

5.2.5 Reciprocity of Relationships

Relationships in Twitter are not reciprocal, a user following another does not imply that the other will choose to follow back. Thus, the existence of reciprocity indicates a stronger relationship between two users as both decided to establish this bond. In the scope of this work, the relationship strength is also viewed in terms of the topical similarity, thus, we expect that reciprocal dyads have a higher similarity than non-reciprocal dyads. This will be verified by both *mentions* and *follow* relationships, i.e., relationships wherein the two users mentioned each other and relationships wherein the two follow each other. We first present the result regarding reciprocal mentions in Figure 5.7. The two distributions differ: the Kolmogorov-Smirnov statistic is of 0.22 with p-value $< 2.2^{-16}$; the median similarity of the distribution of nonreciprocal mentions is 0.08 and of the reciprocal mentions is 0.12. The distribution of similarity for the reciprocal mentions is concentrated in higher values

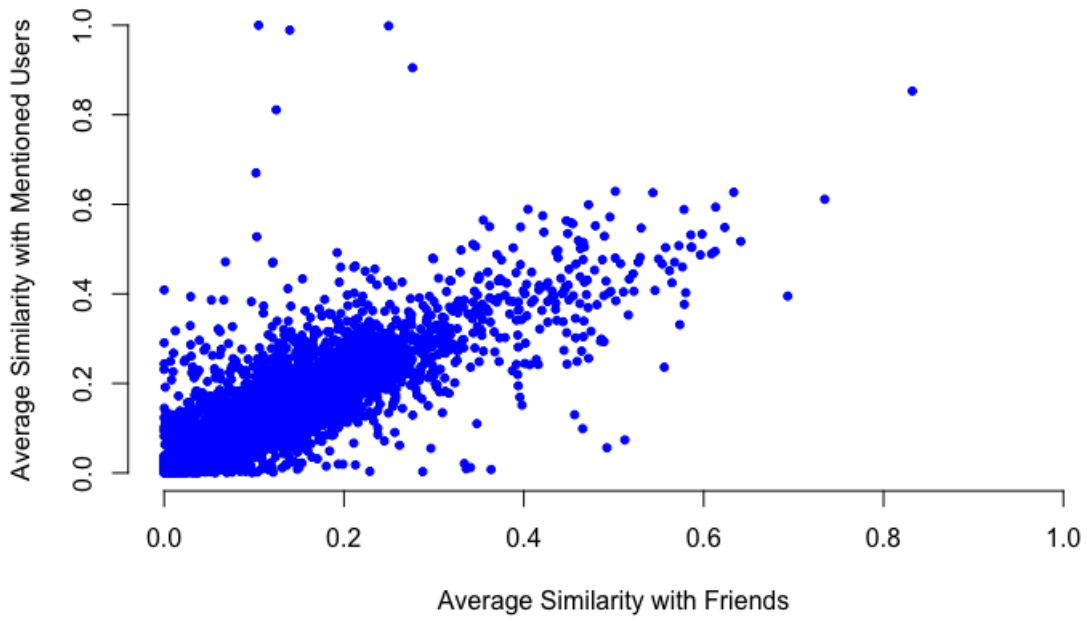


Figure 5.6: Scatter plot wherein each point corresponds to the average similarity between a central user and the users she follows and the average similarity between the central user and the users mentioned by her. . Except for a few outliers the two variables are well correlated.

of similarity, as confirmed by the Mann-Whitney U test with an effect size of 0.64 and a p-value $< 2.2^{-16}$. This indicates that reciprocal relations are more prone to have a higher topical similarity, i.e., users have a more similar topic affiliation if they have a reciprocal relationship.

To obtain the reciprocity of friends connections we used the new connections dataset, i.e., the dyads crawled on February of 2016, which includes data concerning reciprocity of connections. As previously mentioned, Twitter is highly dynamic and it is not accurate to analyze the reciprocity of connections comparing the new friends dyads with the original friends dyads. Thus, we decided to compare the average similarity with reciprocal friends to the average similarity with the nonreciprocal friends from the new connections. The comparison of these two distributions is shown in Figure 5.8. The medians of the distributions of reciprocal friends and of nonreciprocal friends are of 0.12 and 0.07, respectively. As occurred with mentions, the distribution of reciprocal friends is more probable to have a higher value of similarity as it is evidenced by the Mann-Whitney U test, which yielded an effect size of 0.66 with a p-value $< 2.2^{-16}$, furthermore, the Kolmogorov-Smirnov test between the two distributions is of 0.27 with a p-value $< 2.2^{-16}$.

The tests conducted in this subsection reinforce what was seen in Section 5.2.4, i.e., there is no significant difference between the nature of *mention* and *follow* relationships with respect to topical similarity. The distributions of both relationships are very alike when considering the dyads similarity, even with reciprocal relationships. This result can be an important indicator that the motives that make users follow might not be different than the ones that make they mention, at least with respect to topical similarity.

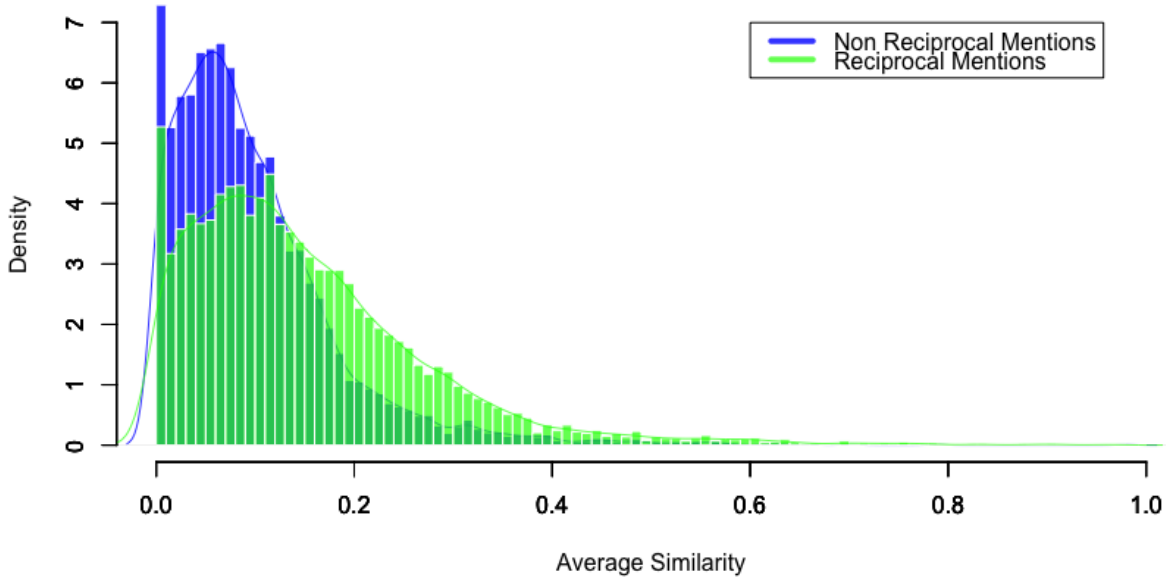


Figure 5.7: The distribution of averages of similarities between central users and users they mentioned and which mentioned them back, as well as the distribution of averages of similarities between central users with users mentioned by them which didn't mention them back. The analysis only concerns the central users that had friends in both situations, a total of $8,663$ users.

Furthermore, we could verify that, in the case of reciprocal relationships, there is a higher inbreeding topical homophily than with nonreciprocal relationships. This indicates that users that stay connected for a sufficiently long period tend to become more similar, by the social influence process, or, conversely, that users similarity can be a factor which influences them to be connected for a longer period.

5.3 Mention Probability

If a central user a follows a user b, there is a dyad involving the two. All the analyses shown until now indicate that the similarity of the dyads is concentrated mostly in low values, relative to the possible range of values $[0, 1]$. Therefore, it is natural to presume that most of the mentions done by the central users involve users which have a low similarity value with them. We confirmed this through Figure 5.9. It shows the probability density function of the similarity involved in the mention, i.e., the similarity between the user that does the mention and the user mentioned. It considers all mentions done by a central user in which the user mentioned were her friend. It can happen multiple times for the same dyad. Specifically, for each mention there is the value of the corresponding dyad similarity in the sample space, e.g., if a user a mentioned friend b twice, which has a similarity of 0.2 with her, the sample space will contain the values $\{0.2, 0.2\}$. As presumably most of the dyads have a low similarity, most of the mentions occur on dyads that have a low similarity. However, this does not imply that users belonging to *follow* dyads of lower similarity have a higher probability of being mentioned. There is just more

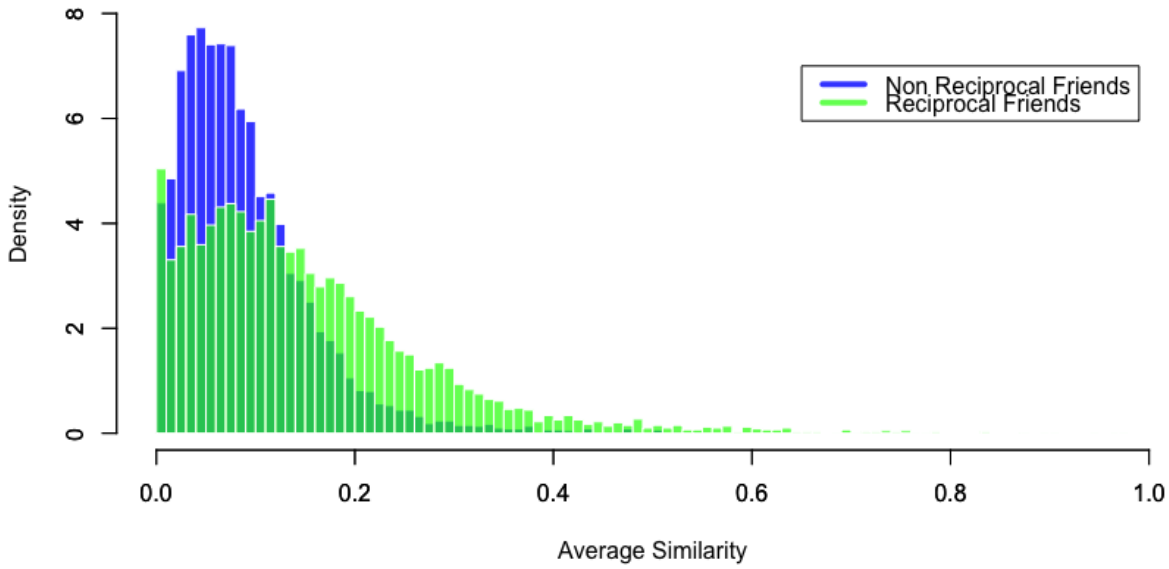


Figure 5.8: The distribution of averages of similarities between central users and users they followed and which followed them back, as well as the distribution of averages of similarities between central users with users followed by them which didn't follow them back. The analysis only concerns the central users that had friends in both situations a total of 5,872 users.

of those dyads. Actually, we expect that users in dyads with high similarity are more likely to be mentioned.

We explored this question, i.e., if the probability of being mentioned is higher for users with a high similarity, by looking at all dyads of friends – i.e., all pair of users wherein one is a central user following another user from the population. We also took into account the number of times that each friend was mentioned. Having the set of all dyads as our sample space, we define two variables, M and S, to verify if the similarity of dyads affected the probability of a user in it being mentioned. As the similarity of the dyads is a continuous variable, we rounded it to two decimal places, producing the discrete variable S. The number of times that the central users mentioned their friends is given by the variable M. Taking $k+1$ as the minimum number of mentions that the dyad must have, the conditional probability of a friend being mentioned by a central user more than k times given the dyad similarity is defined by Equation 5.4.

$$P(M > k | S = s_i) = \frac{P(M > k \cap S = s_i)}{P(S = s_i)} \tag{5.4}$$

Figure 5.10 shows the conditional probabilities of a friend being mentioned by a central user more than $k = 0$, 2, 5 and 10 times, given her similarity with the central user. As expected, the probability decreases when the minimum number of mentions increases. Opposed to the naive interpretation of Figure 5.9, Figure 5.10 shows that friends which have low similarity with the central user do not have a higher probability of being mentioned.

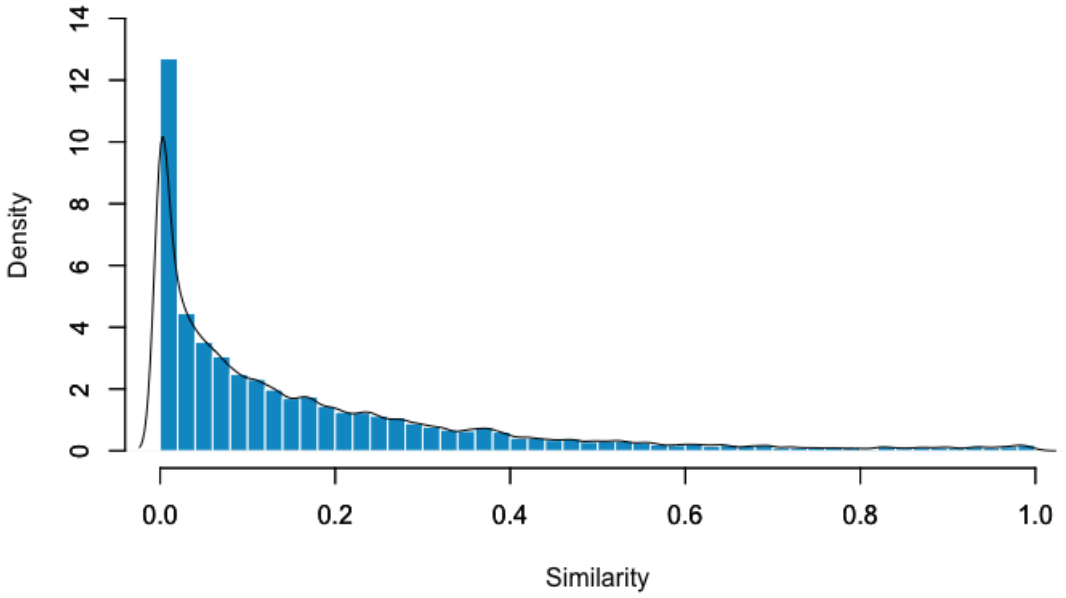


Figure 5.9: Probability density function of similarity of all mentions in which a central user mention one of her friends. A total of 2,010,447 mentions.

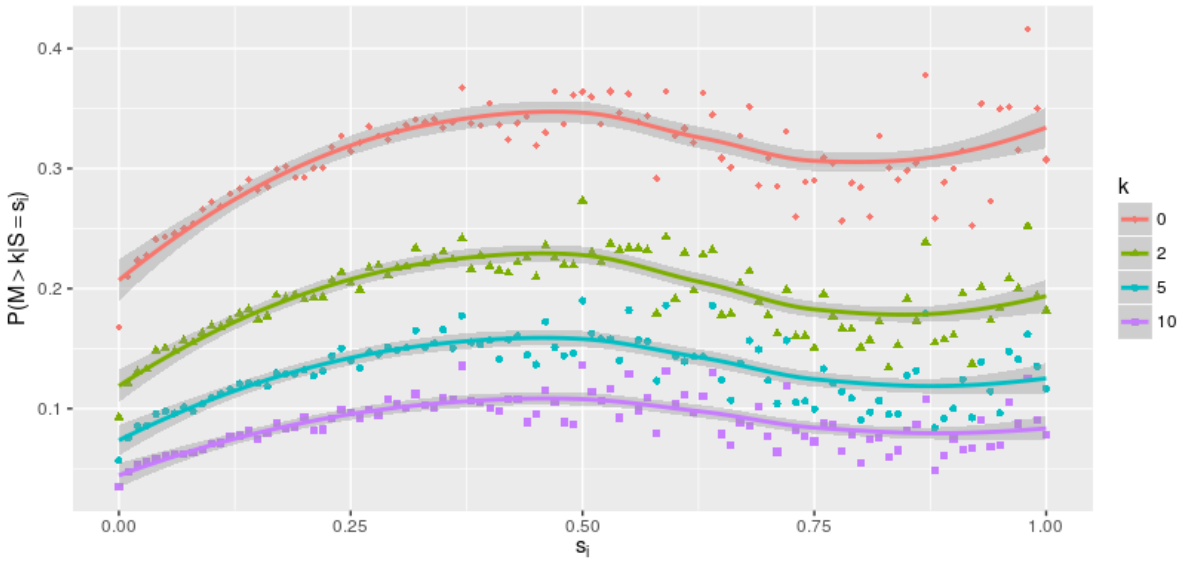


Figure 5.10: The conditional probability of a friend being mentioned more than k times by a central user, given their rounded similarity. Analysis executed with a set of 547,346 dyads.

Actually, users in dyads with a rounded similarity of 1 have a probability of 0.3 of being mentioned at least once, while for dyads with a rounded similarity of 0 the probability is significantly lower, 0.16.

It is observed a stable growth until 0.6 of similarity, after 0.6 there is a non-expected decrease and the data become more disperse. It appears that from that point on, the similarity is not significantly determinant. It remains an open question whether this pattern will persist in future analysis. Overall, the pattern of conditional probabilities appears to be the same for greater values of k, there is only in a reduction in the value of the probability, as being mentioned more times is more challenging.

This analysis shows how the similarity gives an indication of the interactions inside connections, at least for some values of similarity. Furthermore, even considering the loss of stable growth after 0.6, this result can be interpreted as an evidence of homophily inside connections, as connected users with a higher similarity, at least below the saturation point, may have a higher probability to interact with each other.

5.4 Predictor

In Figure 5.6 we verified that there is a correlation between users average similarity with friends and with mentioned users. It indicates that users, on average, follow and mention other users in a similar fashion with respect to topical similarity. Furthermore, the distribution of similarity averages is a right-skewed distribution. However, until now, we did not provide a way to verify if the similarity between users is intrinsically related to their connections as a cause of an effect. Specifically, we did not look if the topical similarity between users can indicate users connections. Here, we evaluate the similarity according to its relative importance, since even a user with a small average similarity with friends might be connected with some who are the most similar to her. Therefore, the topical similarity between users might be an effective variable to estimate their connections.

We approached this issue as a prediction problem. Our question in this section is: Is it possible to predict friends of users only looking at their topical similarity with other users? We tested this question via a predictor that tries to predict users friends from a group of users. Strictly speaking, we tested if the users most similar to the central ones could be predicted as their friends from users pools of different sizes. A pool always contains all the central user friends mixed with other candidates randomly selected. As previously mentioned, there is a difference between users similarity averages, which indicates a different pattern of friends. Thus, we separated the results according to different similarity averages, i.e., we grouped users by their average similarity with their friends. Namely, we rounded users average similarity to one decimal place and grouped them in groups of averages 0, 0.2, 0.4 and 0.6. We did not use groups of average 0.8 and 1 as there was a small number of central users with those averages. The results are shown in Figure 5.11.

We tested the predictor with different pool sizes, using the multiplicative constant k, shown in the x-axis of Figure 5.11. For each k, a pool is created for each user. The size

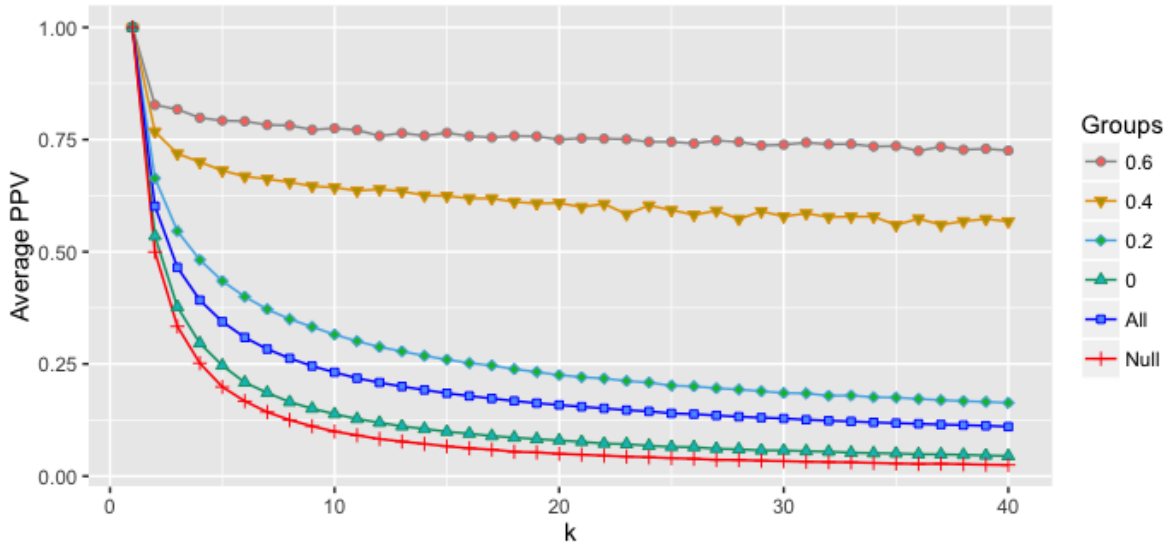


Figure 5.11: Average PPV of the predictor execution for each pool of size $fr(a) \times k$.

of a pool is given by $k \times |fr(u)|$ for each user u, wherein $fr(u)$ is a function that returns the set of friends of user u. The predictor computes the similarity of each central user with the users in the pool and returns a set of predicted friends, which always has size $|fr(u)|$, containing the users that were most similar to the central user. In the y-axis is shown the average PPV(positive predictive values) of the prediction, i.e., the average of the fraction of friends that were correctly predicted. Each line shows the averages for each group of users. The blue line with squares shows the average PPV considering all users together, i.e., users of all averages. The red line with crosses shows the average PPV if the predicted friends were only chosen randomly, also considering all users. For all the groups of users, the prediction was better than only choosing users randomly, indicating that similarity is an elemental feature of users connections.

It is worth mentioning the result observed for users with a average similarity of 0.4 and 0.6. There is a steady PPV as the constant k increases. This can be understood as: even with an increasing set of users to choose from, the predictor keeps correctly returning a significant fraction of their friends and this only happens because they continue to be the most similarly available in the whole pool. This may indicate that there is a local concentration of similarity among some connected users, i.e., the similarities of some central users with their friends have values that are not achieved with other users. We believe that this outcome is created because the individual affiliation pattern in the topics is considerably unique among some connected users, thus the majority of other users in the pools cannot have a greater similarity than the similarity of the friends.

The results regarding all users are not so impressive. Nonetheless, it is important to notice that the method applied here does not take into consideration the social network structure, which is probably the main factor responsible for determining the connections in a network. Our focus is to explore the relation between information and users relationships, not to provide a complete algorithm for link prediction or link recommendation. Nonetheless, we believe that we provide evidence that users affiliation in topics of information can be an important feature to be taken into account in methods interested in the

precision of the prediction.

The results obtained in these last sections corroborate the theories which advocate that relationships among individuals are fundamentally related to their similarity. In this chapter, we were able to show how those theories apply regarding topical similarity in the context of an information network. Besides connected users being more likely to have a higher similarity, the reciprocity of their connections and the probability of their interactions are also related to their affiliation in topics. Moreover, we verified that similarity alone gives a good indication of some users friends, which show how important information is with respect to users relationships. In the next chapter, we further explore the importance of topics in users relations.

Chapter 6

Users' Behavior according to Topics

Until now we have explored how the topical similarity is related to individuals relationships on Twitter. In the previous analyses, topical homophily was based on a single measure composed of the similarity between users vectors representing their affiliation in the topics. We have not analyzed how social relationships are affected by the presence of topics. This chapter focuses on the direct relations between users affiliation in topics and their behavior. We consider that a user is affiliated with a topic if she has tweeted, at least, one hashtag belonging to this topic, i.e., a user, represented by a feature vector $\mathbf{u}$, has to have the value of the feature $u_i > 0$ in order to be affiliated with the topic i. Section 6.1 is concerned with the likelihood of users having a relationship when they share topics; Section 6.2 explores changes of homophily in groups of users affiliated in different topics.

6.1 Homophily on Shared Topics

In this section, we explore if users are more probable to be following or mentioning a user according to the topics they share. We consider that a topic is shared by two users if both are affiliated with it, i.e., both have in their vectors $u_i > 0$ for the topic i. We tested two hypothesis, the first is if a central user is more likely to be following or have mentioned another when they share some topic. It is represented in the Inequalities 6.1 and 6.2. In those inequalities, $P(F_c)$ and $P(M_c)$ stand for the probability that a user is being followed and have been mentioned by the central user c, respectively. $P(T_c)$ is the probability that a user is sharing a topic with the central user c.

$$P(F_c|T_c) > P(F_c) \tag{6.1}$$

$$P(M_c|T_c) > P(M_c) \tag{6.2}$$

Our second hypothesis is if a central user is more likely to have mentioned or to be following another user when this user shares the topic wherein the central user was most active. We consider the most active topic of a user as the topic whereof she had tweeted

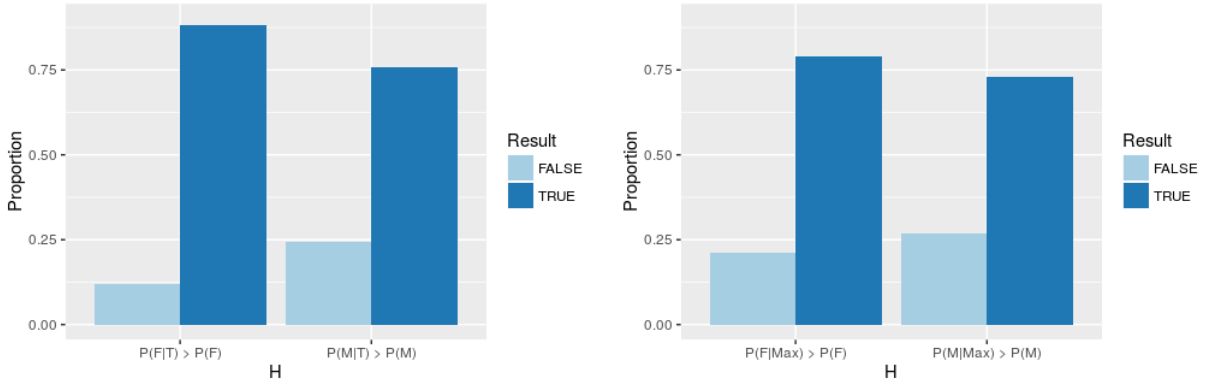


Figure 6.1: Proportion of users wherein each inequality held *true* or *false*.

hashtags more times, considering our whole dataset. The test with respect to following another user is represented in Inequality 6.3 and with respect to mentioning in Inequality 6.4. In both tests, $P(Max_c)$ is the probability that a user shares the topic wherein the central user c has been most active.

$$P(F_c|Max_c) > P(F_c) \tag{6.3}$$

$$P(M_c|Max_c) > P(M_c) \tag{6.4}$$

The calculation of the conditional probabilities are straightforward. Here we only demonstrate in Equation 6.5 the calculation of $P(F_c|T_c)$, which is required for Inequality 6.1.

$$P(F_c|T_c) = \frac{P(F_c \cap T_c)}{P(T_c)} \tag{6.5}$$

In our dataset, there are users which have not followed and mentioned any user in some of the presented probabilities. For instance, a central user c might not have followed someone in the topic wherein she has been most active. In this case, $P(F_c \cap Max_c) = 0$, which leads to $P(F_c|Max_c) = 0$. These cases are also considered in our analysis.

The number of times that each inequality was *true* or *false* is shown in Figure 6.1. It shows that for the high majority of central users all the inequalities hold true. It indicates that most users are more likely to be following and have mentioned users that share some topic with them. We expected that the fraction would be larger for users that share the topic wherein the central user was more active, however, the proportion is smaller. Though, as shows Figure 6.2, the mean probability of have been mentioned and are being followed in this case are notably larger than all the other cases.

This result contributes to the idea that individuals tend to have a dyad with the ones that they have interests in common. Differently from the previous chapter, this analysis verified the presence of homophily when two users share a characteristic, in this case, a topic. The results hold for *follow* and *mention* relations and indicate that, either users

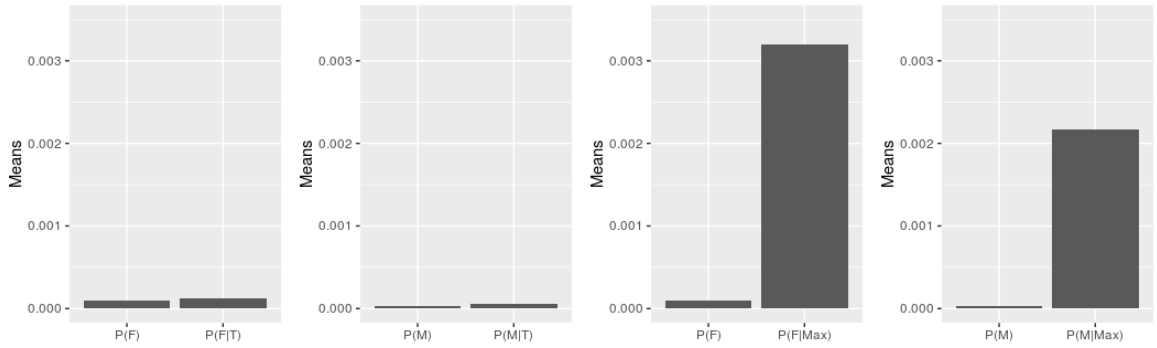


Figure 6.2: Means the probabilities of a user being followed and mentioned.

tend to select their relationships according to topics or their topic affiliations are influenced by their relationships. However, it is important to notice a limitation of these results. As shown in Figure 6.2, despite the means of the conditional probabilities been higher in all cases, the values of all probabilities are very low. This is likely caused by a large number of users in all samples, which led to the fraction of users followed and mentioned by any user being small all cases. Nonetheless, their values imply that these results cannot be used alone to indicate the mechanisms of users relationships.

6.2 Behavior in topics

In the previous analysis, we have shown that the sharing of topics is related to users relationships. However, we were not concerned about how users behave differently according to the topics they are affiliated. Our hypothesis is that users inserted in different topics behave in different ways regarding their relationships. Furthermore, if this hypothesis is confirmed, is interesting to know what could be the reasonings behind this process. We analyzed if this difference exists by two approaches: firstly, we estimated the level of inbreeding homophily considering only the users affiliated in each topic and, secondly, we looked at the probabilities of interaction inside different topics. These two approaches are detailed in the following subsections.

6.2.1 Different Levels of Inbreeding Homophily

In this work, we detect the inbreeding homophily by assessing the difference between the average similarity with friends and the average similarity with random users, as was done in Section 5.2. Here we tested if there were significant differences between the inbreeding homophily of topics. Strictly speaking, considering only the users affiliated with each topic, we calculated the difference between the distribution of average similarities of central users with friends and with randomly chosen users. As detailed in Section 5.2.1, we assume that the level of inbreeding homophily can be measured through the Kolmogorov-Smirnov statistic.

To assure that we had sufficient statistics, we computed the Kolmogorov-Smirnov statistic for each topic that had more than $1,000$ users in the population and more than 30

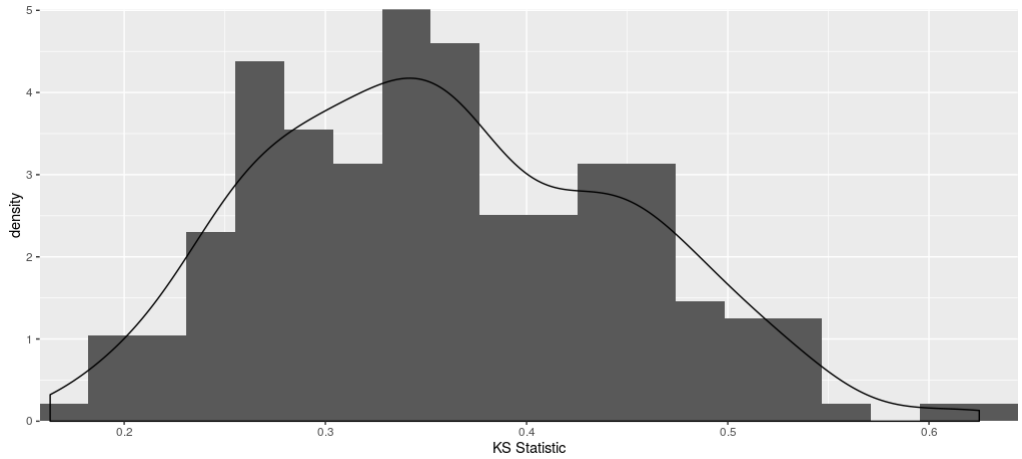


Figure 6.3: Histogram of the values of the Kolmogorov-Smirnov statistic between the distribution with friends and the one with random users.

Figure 6.4: Topics wherein the KS statistic was bigger.

central users. Furthermore, we did not consider the topics in which the KS test had a p-value > 0.001, to assure that the distributions did not only differ by chance.

The statistics for these topics are summarized on Figure 6.3. The level of inbreeding homophily is distributed in the ranges between 0 and 0.6. We manually verified some of the topics wherein the test has yielded the highest and lowest values and, in a first glance, it appears that exists some pattern. We show in Figure 6.4 two topics that had a high value of the KS statistic, 0.52 and 0.53. These topics contain hashtags related to sports competitions and it is possible to believe that the distributions differ as there might be opposing groups of users which have interest in the competition, but differing in the majority of other aspects. On the other hand, we show in Figure 6.5 topics with a low KS statistic, 0.21 and 0.19, which indicates that the distributions are more similar. The hypothesis that emerges from these images is that some sets of hashtags may be part of a dialect used by a group of users that are highly homogeneous among them. This analysis may guide questions for future research, however, they are outside of the scope of this work.

6.2.2 Mentions in Topics

We have shown that there are different values of inbreeding homophily according to the topics. In this subsection, we focus on users interactions inside topics. Strictly speaking,

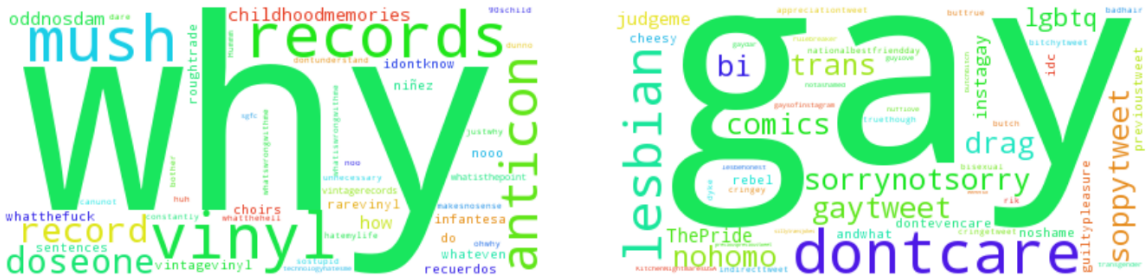


Figure 6.5: Topics wherein the KS statistic was smaller.

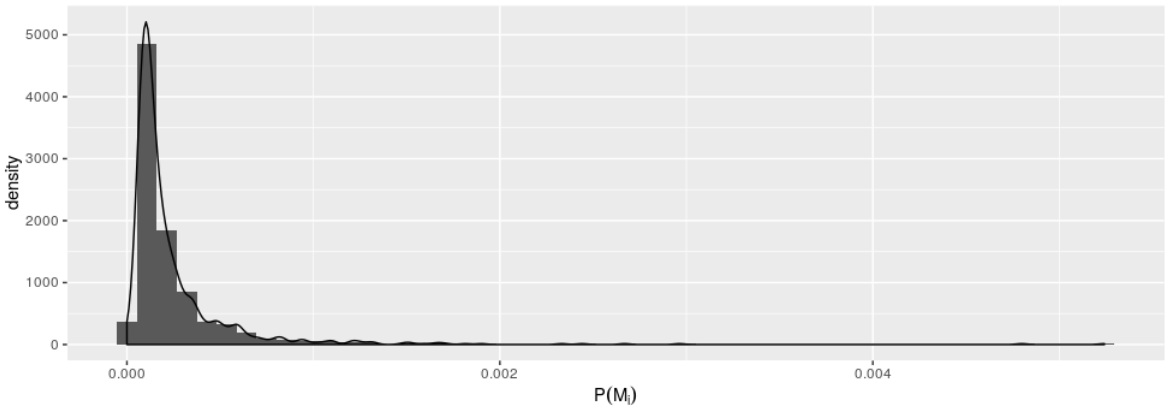


Figure 6.6: Histogram of probabilities of a user inside a topic being mentioned by any central user.

we are concerned with the probabilities of a user being mentioned inside each topic. We define $P(M_i)$ as the probability of a user being mentioned by any central user, with respect to all the users affiliated with the topic i. Figure 6.6 shows the histogram of probabilities calculated for those topics having at least 30 central users affiliated. Our first idea was to apply a similar procedure of Section 6.2.1, i.e., to verify if could be a semantic interpretation behind the probability values. However, in a manual verification, we did not find any interpretation.

We then tried to see if the result could be a consequence of a more fundamental variable, which could be correlated with the number of users affiliated with the topic. Figure 6.7 shows a log-log scatter plot of the probability by the number of users. It indicates that the probability of a user being mentioned has an inverse power relationship with the size of the topic.

On the other hand, Figure 6.8 shows that there is not a reduction in the numbers of mentions of the topics as their numbers of users grow. The number of mentions of a topic is the number of mentions done by a central user, such that she and the mentioned user were affiliated with the topic. We interpret the probability decrease as a possible result of two phenomena not necessarily unrelated. The first supposes that smaller topics have a higher concentration of people alike. Thus, their users might be closer and, by the homophily principle, might be more prone to interact than users from bigger topics, which contain a broader variety of users. The second interpretation is based on the preferential attachment effect. As the size of the group increases, the mentions tend to be more concentrated around a smaller number of users with high degree, reducing the

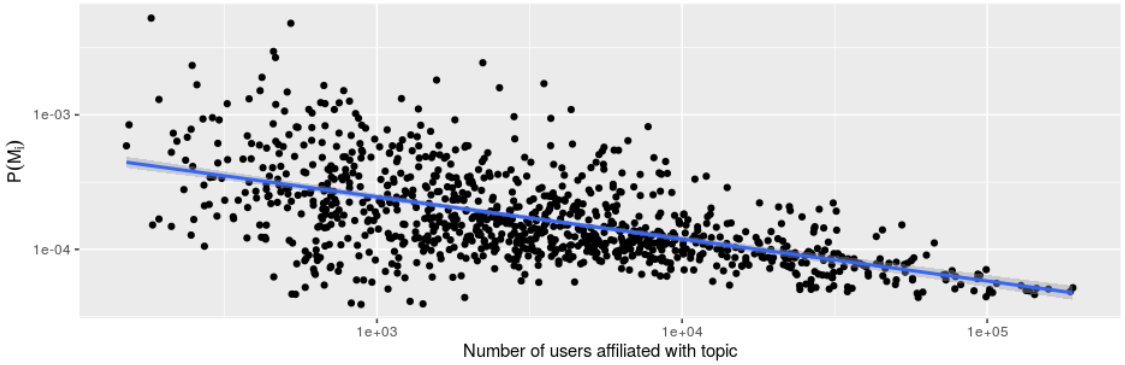


Figure 6.7: Probability of a user being mentioned in each topic by the number of users in it. A log-log regression of the two variables gives $\alpha = 0.0021$ and $\beta = -0.31$ for a function $y = \alpha x^{\beta}$.

total fraction of users mentioned. Further analysis might give a thorough explanation of this phenomena.

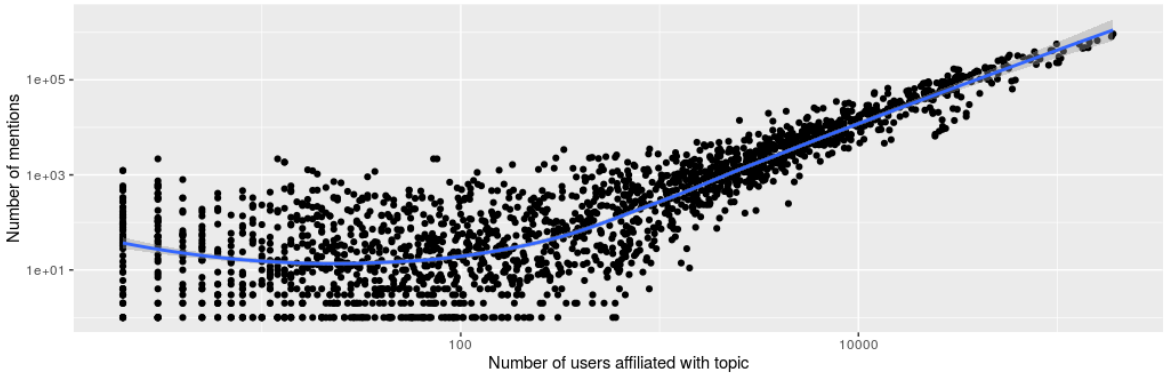


Figure 6.8: The total number of mentions in each topic by the number of users in it.

This result demonstrates that, despite homophily being significant in people relationships, other properties and the arrangement of users relationships can affect their behavior. This indicates the necessity of more research in homophily taking into account the social network structure.

Chapter 7

Conclusion

In today's world, online social networks as Twitter provide a laboratory where information and users connections are available for study. This dissertation is concerned with how the structure of a social network is related to the information shared in it. The connections in a social network are the substrate over which information flows, which makes their flow partially dictated by the network structure. However, information flow cannot be seen as an independent phenomenon; it encapsulates contents that can affect how individuals behave. For instance, people might be inclined to bound to others according to their affinity concerning the information they share. On the other hand, depending on the information one advocates, some might prefer to not bond with her. We explored this relation using the Twitter data and we found that individuals which have a relationship tend to be similar regarding the information they share.

One of the premises to investigate how information is coupled with social connections is to design a model which captures its desired characteristics. We achieve this by modeling information as semantic topics of hashtags as Weng et al. [74]. These topics encompass contents of information shared among users. We computed users participation in topics to characterize individuals interests and preferences on Twitter. This characterization served as a basis for the exploration of topical similarity between individuals and we found that, on average, individuals tend to have a relationship with users more similar to them than with a random group of users. The dyads of some users might experience a greater influence of the topical similarity. This is so profound to some users, that they are essentially connected to the users most similar them in the network, what suggests an effective way to predict friends, at least, for some users.

We verified if the influence of the topical similarity between individuals differed in *mentions* and *follows* relations. Our results show a consistency across the two types of relationships, showing no significant difference between them. This was also verified when considering reciprocal relationships, which, in both cases, showed a higher level similarity than in a non-reciprocal relationship.

Modeling users according to their affiliation in topics also allowed us to verify how their adoption of topics is associated with their relationships. Our results show that groups of users affiliated with different topics show different levels of inbreeding homophily and we suggest that this can be caused by the topic nature. Groups of users in different topics also mention each other with different probabilities, what appears to be a result of the

size of the topic and the preferential attachment effect. Furthermore, we verified that the majority of users are more likely to mention others that share topics with them than the others who don't share.

The approach presented in this work uses hashtags to build the topics of information. It limited our results to only users that used hashtags, what significantly reduced our sample. Moreover, as we did not have the whole Twitter network structure, our hypothesis was restricted to a set of dyads and could not explore questions involving network measures such as distance and centrality. Nonetheless, we believe that our sample provides a significant support to understand some relationships among users. Another limitation is our method to build topics, which ignores the temporal behavior of hashtags. The moment in which the hashtags co-occur might contain specificities that we are not able to capture. However, even with these limitations, we could verify that the topics detected have a semantic sense and our set of users were sufficient to achieve relevant results.

Our work provides evidence of a greater topical similarity between connected individuals, which may be seen as an evidence of inbreeding homophily, as defined by McPherson & Smith-Lovin [49]. We also deepen the understanding of how the information that traverses individuals connections can affect their behavior.

This is a significant achievement involving social hypothesis using Twitter Data, but our contributions include providing a feasible computational way to compute the similarity between users and assess homophily in a social network. This can be further enhanced to improve the understanding of the mechanisms by which users are connected analyzing the whole social network structure, which was not available in our work. Furthermore, it is necessary further investigation of how the flow of information is related to the network dynamics. The results that we obtained with topics also leave open opportunities to explore how their semantics affect the behaviors of users who adopt them. Other possibilities lie in using our methods in applications for friendship recommendation or finding missing links in a social network.

Bibliography

[1] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–4, aug 2010.

[2] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2):1–33, may 2012.

[3] R. Axelrod. The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution*, 41(2):203–226, apr 1997.

[4] Lars Backstrom and Jure Leskovec. Supervised Random Walks: Predicting and Recommending Links in Social Networks. *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, page 635, feb 2010.

[5] Yaneer Bar-Yam. *Dynamics of Complex Systems*, volume 12. Westview Press, 1997.

[6] A. Barabási. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, oct 1999.

[7] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*, volume 1. Cambridge University Press, Cambridge, 1st edition, may 2008.

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

[9] Stefano Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, feb 2006.

[10] Bjoern Bringmann, Michele Berlingerio, Francesco Bonchi, and Arisitdes Gionis. Learning and Predicting the Evolution of Social Networks. *IEEE Intelligent Systems*, 25(4):26–35, jul 2010.

[11] Björn Bringmann, Siegfried Nijssen, and Albrecht Zimmermann. Pattern-Based Classification: A Unifying Perspective. In *Proc. LeGo From Local Patterns to Global Models Second ECML PKDD Workshop*, nov 2011.

[12] Camila Buono, Lucila G Alvarez-Zuzek, Pablo a. Macri, and Lidia a. Braunstein. Epidemics in Partially Overlapped Multiplex Networks. *PLoS ONE*, 9(3):e92200, mar 2014.

[13] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E*, 74(3):036116, sep 2006.

[14] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, may 2009.

[15] D. Centola, J. C. Gonzalez-Avella, V. M. Eguiluz, and M. San Miguel. Homophily, Cultural Drift, and the Co-Evolution of Cultural Groups. *Journal of Conflict Resolution*, 51(6):905–929, dec 2007.

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

[17] Damon Centola and Michael Macy. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 113(3):702–734, nov 2007.

[18] Nicholas A. Christakis and James H. Fowler. Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior. *Statistics in Medicine*, pages 1–32, sep 2011.

[19] Valerio Ciotti, Moreno Bonaventura, Vincenzo Nicosia, Pietro Panzarasa, and Vito Latora. Homophily and missing links in citation networks. *EPJ Data Science*, 5(1):7, dec 2016.

[20] Aaron Clauset, Cristopher Moore, and M E J Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, may 2008.

[21] W. J. Conover. *Practical nonparametric statistics*. Wiley, 3rd edition, 1999.

[22] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, jul 2012.

[23] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proc. KDD'08*, page 160, New York, New York, USA, 2008. ACM Press.

[24] P Erdős and A Rényi. On the Evolution of Random Graphs. In *Publication of The Mathematical Institute of The Hungarian Academy of Sciences*, pages 17–61, 1960.

[25] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. *ACM SIGCOMM Computer Communication Review*, 29(4):251–262, oct 1999.

[26] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, feb 2010.

[27] Santo Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, jan 2007.

[28] Daniel Foxman and Gregory Bateson. Steps to an Ecology of Mind. *The Western Political Quarterly*, 26(2):345, jun 1973.

[29] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE*, 6(8):e22656, aug 2011.

[30] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1:197, jan 2011.

[31] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27, jul 2014.

[32] MS Granovetter. The Strength of Weak Ties. *American journal of sociology*, 76(6):21, 1973.

[33] Steve Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, oct 2010.

[34] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. WTF: the who to follow service at Twitter. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 505–514, New York, New York, USA, may 2013. ACM Press.

[35] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *Proceedings of the SDM Workshop on Link Analysis, Counterterrorism and Security*, 2006.

[36] Paul Heymann and Hector Garcia-Molina. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. *Stanford InfoLab Technical Report*, (2006-10), 2006.

[37] B. W. Kernighan and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. *Bell System Technical Journal*, 49(2):291–307, feb 1970.

[38] Gueorgi Kossinets and Duncan J. Watts. Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, 115(2):405–450, sep 2009.

[39] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato. Finding Statistically Significant Communities in Networks. *PLoS ONE*, 6(4):e18961, apr 2011.

[40] David Laniado, Yana Volkovich, Karolin Kappler, and Andreas Kaltenbrunner. Gender homophily in online dyadic and triadic relationships. *EPJ Data Science*, 5(1):19, dec 2016.

[41] T. Winograd Larry Page, Sergey Brin, R. Motwani. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

[42] P.F. Lazarsfeld and R.K. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 1954.

[43] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter Trending Topic Classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE, dec 2011.

[44] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web - WWW '12*, page 251, New York, New York, USA, apr 2012. ACM Press.

[45] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic Evolution of Social Networks. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 462, aug 2008.

[46] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, New York, USA, nov 2003. ACM Press.

[47] Kenneth O. McGraw and S. P. Wong. A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365, 1992.

[48] J. Miller McPherson and Lynn Smith-Lovin. Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups. *American Sociological Review*, 52(3):370, jun 1987.

[49] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, aug 2001.

[50] R Michalski, P Kazienko, and D Krol. Predicting Social Network Measures Using Machine Learning Approach. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1056–1059. IEEE, aug 2012.

[51] JL Moreno. Emotions mapped by new geography. *New York Times*, 1933.

[52] S A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information Network or Social Network? The Structure of the Twitter Follow Graph. *WWW'14 Companion*, pages 493–498, apr 2014.

[53] Seth A. Myers and Jure Leskovec. Clash of the Contagions: Cooperation and Competition in Information Diffusion. In *2012 IEEE 12th International Conference on Data Mining*, pages 539–548. IEEE, dec 2012.

[54] Seth A. Myers and Jure Leskovec. The bursty dynamics of the Twitter information network. In *Proceedings of the 23rd international conference on World wide web - WWW '14*, pages 913–924, New York, New York, USA, apr 2014. ACM Press.

[55] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, jan 2003.

[56] M E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, jun 2006.

[57] M. E. J. Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31, dec 2011.

[58] Siegfried Nijssen and Joost N. Kok. The Gaston Tool for Frequent Subgraph Mining. *Electronic Notes in Theoretical Computer Science*, 127(1):77–87, mar 2005.

[59] Jonathan A. Obar and Steve Wildman. Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy*, 39(9):745–750, oct 2015.

[60] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, 86(14):3200–3203, apr 2001.

[61] Matjaž Perc. The Matthew effect in empirical data. *Journal of the Royal Society, Interface / the Royal Society*, 11(98):20140378, jul 2014.

[62] Evelyn Perez-Cervantes, Jesus P. Mena-Chalco, Maria Cristina F. De Oliveira, and Roberto Cesar. Using Link Prediction to Estimate the Collaborative Influence of Researchers. In *2013 IEEE 9th International Conference on e-Science*, pages 293–300. IEEE, oct 2013.

[63] Prabhakar Raghavan. It's time to scale the science in the social sciences. *Big Data & Society*, 1(1):2053951714532240, apr 2014.

[64] John A. Rice. *Mathematical Statistics and Data Analysis*. Cengage Learning, 3 edition, 2006.

[65] Dawn T Robinson, Laura Aikens, and Laura Aikens. Homophily. In *Encyclopedia of Group Processes {{}{&}{}} Intergroup Relations*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States, 2009.

[66] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics. In *Proceedings of the 20th international conference on World wide web - WWW '11*, page 695, New York, New York, USA, mar 2011. ACM Press.

[67] Giulio Rossetti, Michele Berlingerio, and Fosca Giannotti. Scalable Link Prediction on Multidimensional Networks. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 979–986. IEEE, dec 2011.

[68] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–23, jan 2008.

[69] David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 4 edition, jan 2007.

[70] M. V. Tamm, A. B. Shkarin, V. A. Avetisov, O. V. Valba, and S. K. Nechaev. Islands of Stability in Motif Distributions of Random Networks. *Physical Review Letters*, 113(9):1–5, jun 2014.

[71] Hanghang Tong, Christos Faloutsos, and Jia-yu Pan. Fast Random Walk with Restart and Its Applications. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 613–622. IEEE, dec 2006.

[72] Peter D. Turney and Patrick Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, jan 2010.

[73] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[74] Lilian Weng and Filippo Menczer. Topicality and Impact in Social Media: Diverse Messages, Focused Messengers. *PLOS ONE*, 10(2):e0118410, feb 2015.

[75] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Predicting Successful Memes using Network and Community Structure. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)*, page 10, mar 2014.

[76] Lilian Weng, Jacob Ratkiewicz, Nicola Perra, Bruno Gonçalves, Carlos Castillo, Francesco Bonchi, Rossano Schifanella, Filippo Menczer, and Alessandro Flammini. The Role of Information Diffusion in the Evolution of Social Networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, page 356, New York, New York, USA, aug 2013. ACM Press.

[77] Xifeng Yan and Jiawei Han. gSpan: graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 721–724. IEEE Comput. Soc, 2002.

Appendix A

Statistics

Some basic concepts are presented in this appendix for this dissertation to be more self-contained. A more detailed explanation may be found in the references.

A.1 Probability Density Function

A probability density function, $f(x)$, of a continous variable X is defined such that $f(x) \geq 0$ for an event $x \in X$ and $\int_{\infty}^{-\infty} f(x)dx = 1$. The probability of an event on the interval $[a, b]$ is given by:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

Some histograms in this work are shown in density scale and their bins have the same width. Thus, the probability of an outcome is given by:

$$P(x) = f(x) \times width$$

For more information on probability density function, see [64].

A.2 Kolmogorov-Smirnov Test

The Two-sample Kolmogorov-Smirnov samples test is a nonparametric statistical test used for comparing two continuous probability distributions. As it is a nonparametric test, it does not assume anything about the distributions, e.g., if the data are normally distributed. The test check if two samples come from the same distribution looking at their empirical distribution function. It returns a statistic D which is the supremum difference between the two empirical distribution functions at any given point. See more on Conover [21] and Sheskin [69].

A.3 Mann-Whitney U Test

The Mann-Whitney U Test, also known as Wilcoxon rank sum test is a non-parametric test analog to the t-test (which assumes normality of the distributions). It assumes that the data of the distributions are in an ordinal format, i.e., their values represent a score of the observations on an arbitrary scale. It is often used to compare if two samples, which do not follow a normal distribution, come from the same population or if one of them is more likely to have higher scores than the other [69].

When comparing two distributions of size m and n, firstly, the values of the two distributions are joined and put in rank order, i.e., the smallest values will have rank 1, the second smallest rank 2, ..., the highest will have rank $m + n$. After that, each pair m_i, n_j is compared to see which is the higher, resulting on $m \times n$ comparisons. The test has as result the statistic U, which is the number of favorable pairs that supports the hypothesis. The hypothesis cannot be rejected when the number of favorable pairs is higher than the number of unfavorable pairs. There are other mathematically equivalent ways to compute the test statistic U. We preferred this approach as we believe it is the most intuitive. Detailed definitions can be found on Rice [64] and Conover [21].

In the Mann-Whitney U tests done in this work, we have chosen to use the common language effect size [47] as an indicator of the strength of the result. The common language effect size is the proportion of favorable pairs, U, over all pairs. Thus, if the effect size is 0.5, the result is insignificant, as half of the comparisons are favorable to the hypothesis and half are unfavorable. As far as the effect size is further away from 0.5, more impactful is the result.

Appendix B

Supplementary Information

B.1 Dataset Statistics

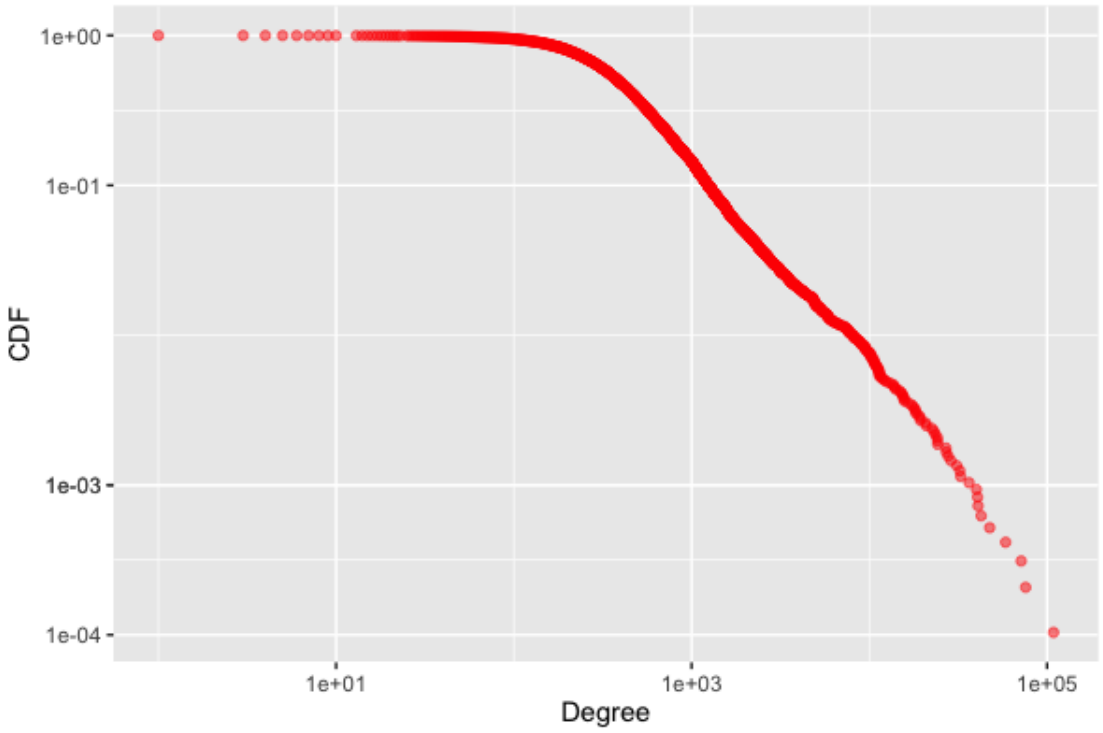
We crawled the follow connections of users in 2013 and 2016. This process had as focus the users that were followed by a set of users denominated central users. In Figure B.1a and Figure B.1b we show the out-degrees of the central users and the in-degrees of all users that are followed by them. In 2016 we also crawled the connections of the users followed by the central users. Figure B.2a shows the out-degree and Figure B.2b shows the in-degree of all users. It is visible that all distributions follow a power law like degree distribution.

B.2 Community Detection

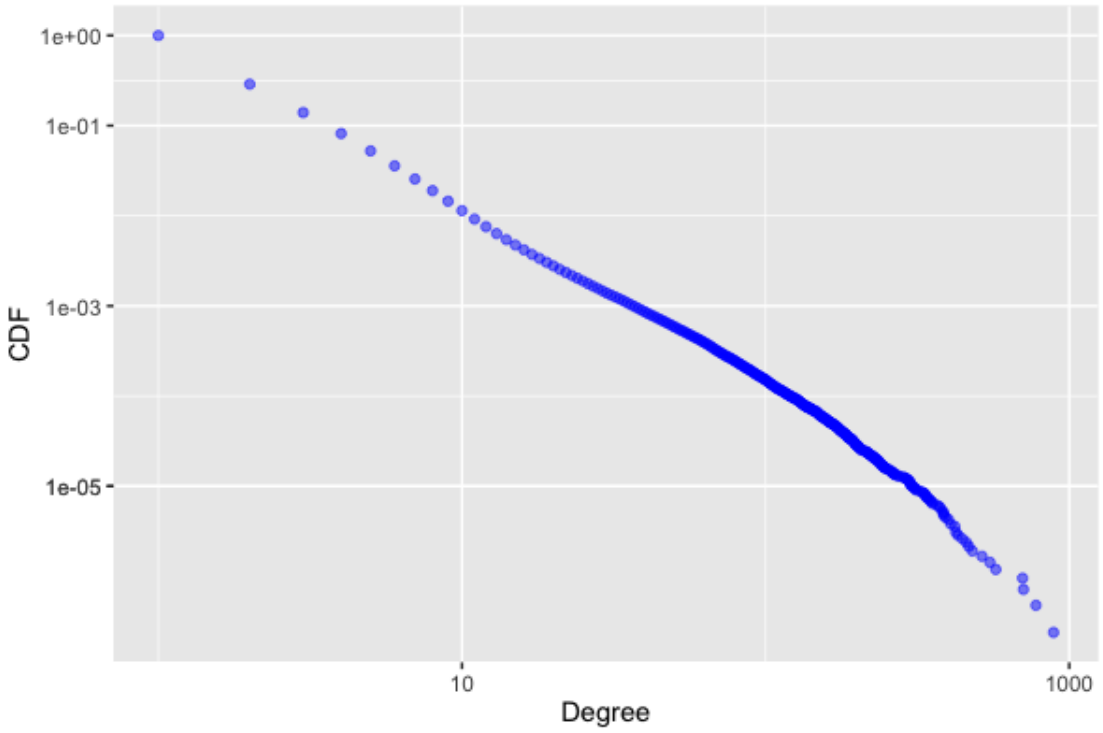
The topics of information used in our analysis are clusters of hashtags. These clusters are built through the OSLOM community detection tool [39]. One of the reasons for the use of OSLOM is its efficiency given that the hashtags co-occurrence graph is relatively big. Its computational complexity is hard to be assessed and has not been specified by the authors. However, they their tests on artificial benchmarks have shown that the time for the execution of the method scales almost linearly with the graph number of nodes.

OSLOM can detect a hierarchical structure of the clusters and, as it is an optimization algorithm, it takes as parameters the number of runs for the first and for the other hierarchical levels. We executed 100 runs in all levels. Furthermore, OSLOM can take as input parameters other efficient community detection algorithms that are executed priorly. The partitions found by them are used as initial conditions. The tool then assess and chooses the best-found modules. Our execution used 10 runs of the Infomap [68], Copra [33] and Louvain [8] methods. The total time of execution with our considered dataset was of approximately one day.

We adopted the first level in the dendogram of communities for our analyses. One could argue that it would be a good choice to use a higher level in this dendrogram, i.e., a higher hierarchical level, as it would reduce the number of communities. This would be good for our analysis if there was a better semantic mapping in a higher hierarchical level. However, in our executions the higher hierarchical levels just group together the

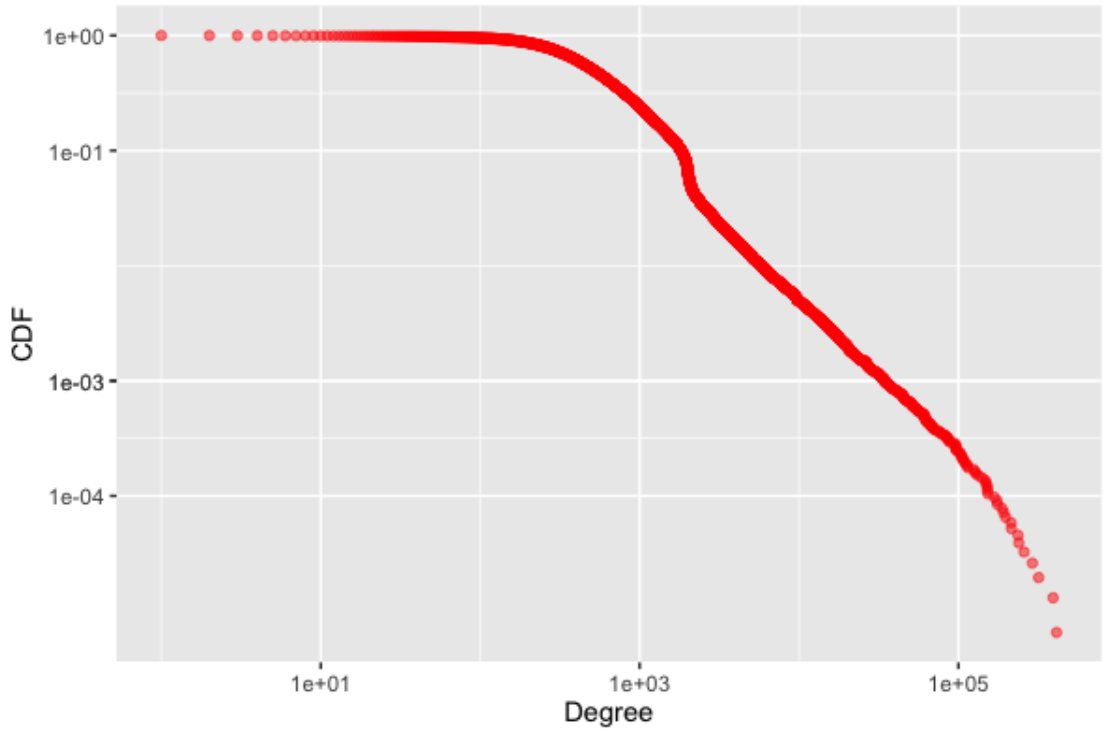


(a) Out-degree distribution of the central users .

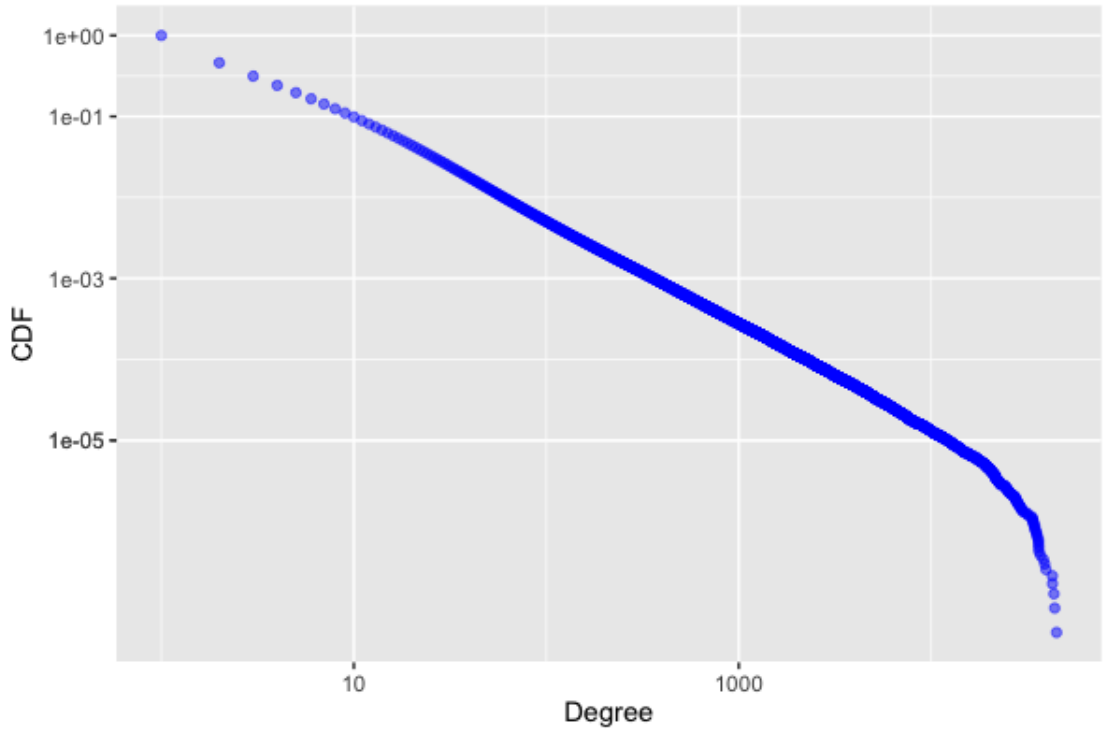


(b) In-degree distribution of all users.

Figure B.1: In/out-degree distributions considering the first crawling of follow connections.



(a) Out-degree distribution of all users.



(b) In-degree distribution of all users.

Figure B.2: In/out-degree distributions considering the second crawling of follow connections.

biggest communities, resulting in a more skewed distribution of the communities sizes. We believe that this worsens the quality of the topics of information.