Universidade Estadual de Campinas
Instituto de Computação

Kleber Andrade Oliveira

Topical Homophily on the Meme Spreading in Online Social Networks

Homofilia por Tópicos no Espalhamento de Memes em Redes Sociais Online

CAMPINAS
2018

Kleber Andrade Oliveira

Topical Homophily on the Meme Spreading in Online Social Networks

Homofilia por Tópicos no Espalhamento de Memes em Redes Sociais 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 Kleber Andrade Oliveira e orientada pelo Prof. Dr. André Santanchè.

CAMPINAS
2018

Agência(s) de fomento e nº(s) de processo(s): CNPq, 131090/2017-8

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Márcia Pillon D'Aloia - CRB 8/5180

OL4t

Oliveira, Kleber Andrade, 1990-
Topical homophily on the meme spreading in online social networks / Kleber Andrade Oliveira. – Campinas, SP : [s.n.], 2018.

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

1. Redes sociais on-line. 2. Ciência social computacional. 3. Sistemas complexos. 4. Dinâmica não-linear. 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 no espalhamento de memes em redes sociais online
Palavras-chave em inglês:
Online social networks
Computational social science
Complex systems
Nonlinear dynamics
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
André Santanchè [Orientador]
Christian Rodolfo Esteve Rothenberg
Julio Cesar dos Reis
Data de defesa: 06-08-2018
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Kleber Andrade Oliveira

Topical Homophily on the Meme Spreading in Online Social Networks

Homofilia por Tópicos no Espalhamento de Memes em Redes Sociais Online

Banca Examinadora:

- Prof. Dr. André Santanchè
  Instituto de Computação - Universidade Estadual de Campinas
- Prof. Dr. Christian Rodolfo Esteve Rothenberg
  Faculdade de Engenharia Elétrica e de Computação - Universidade Estadual de Campinas
- Prof. Dr. Julio Cesar dos Reis
  Instituto de Computação - Universidade Estadual de Campinas

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

Campinas, 06 de agosto de 2018

Acknowledgements

I would like to thank first the person who enabled the pursuing of my scientific career, my mother Maria. Her unfailing support paved the way to everything I achieved and her strong character made me a resilient and focused person.

I am really grateful to my advisor André. His immense patience is what makes him so fit to coordinate a remarkable laboratory with deep roots in interdisciplinarity. This culture was the key ingredient to the birth of a work in contemporary challenges. Felipe, a fine product from this lab and a scientific reference to me, provided me with fundamental guidance through the obstacles of dealing with real-world data.

Everyone at the Laboratory of Information Systems (LIS) had a developmental role in my views about science and how it should be done. Fagner, Jacqueline, Matheus, Ivelize, Victor and Leandro to name a few of them. I would like to thank also people at the Cosnet, in the Universidad de Zaragoza, for the brief but intense period of research which enriched this work. Professor Yamir, Alberto, Sergio, Aga, Carlos and Javi all helped me go through this amazing experience outside of my country.

My other friends Rafael, David, Marina and Guilherme, who are also colleagues, contributed to reduce my ever-present feeling of being a fish out of water, as I sometime delve into areas which I am not familiar with.

I shall never forget Monica, a fantastic human being who nurtured me at every turnaround in the many obstacles life had to me in this period.

My second family who made my time in Campinas packed with the sweetest memories one could desire. Gabriel, Janaina, Victor, Danilo, Leandro, Oscar, Leo, Eduardo and many others from our great group. Thank you very much!

Resumo

Um dos problemas centrais na ciência social computacional é entender como a informação se espalha em redes sociais online. Alguns trabalhos afirmam que pessoas que usam estas redes podem não ser capazes de lidar com a quantidade de informação devido às restrições cognitivas, o que resulta em um limite de atenção gasta para ler e compartilhar mensagens. Disso emerge um cenário de competição, em que memes das mensagens visam ser lembrados e compartilhados para que durem mais do que os outros. Esta pesquisa está preocupada em construir uma evidência empírica de que a homofilia desempenha um papel no sucesso de cada meme na competição. A homofilia é um efeito observado quando pessoas preferem interagir com aqueles com os quais se identificam. Coletando dados no Twitter, nós aglomeramos memes em tópicos que são usados para a caracterização da homofilia. Executamos um experimento computacional, baseado num modelo simplificado de memória para adoção de memes, e verificamos que a adoção é influenciada pela homofilia por tópicos.

Abstract

One of the central problems in the computational social science is to understand how information spreads in online social networks. Some works state that people using these networks may not cope with the amount of information due to cognitive restrictions, resulting in a limit of attention spent reading and sharing messages. A competition scenario emerges, where memes of messages want to be remembered and shared in order to outlast others. This research is concerned with building empirical evidence that homophily plays a role in the success of each meme over the competition. Homophily is an effect observed when people prefer to interact with those they identify with. By gathering data from Twitter, we clustered memes into topics that are used to characterize the homophily. We executed a computational experiment, based on a simplified memory model of meme adoption, and verified that the adoption is influenced by topical homophily.

List of Figures

2.1 Difference between broadcast (left) and viral diffusion (right). Edges represent meme adoptions. Source: [30]. 19

2.2 Each user has two discrete memory slots. At a given time, the shaded user may create a new meme (m_9) with probability μ or reshare an old one in his memory (m_6) with probability $1 - \mu$, and his followers may adopt it by replacing one of the memes in their memories. Source: [51]. 20

2.3 Examples of different topic clusters and how they are connected after being retrieved from the hashtag co-occurrence network. The size of each circumference indicates how many hashtags are inside the cluster. Source: [61]. 24

3.1 Adoption of the meme h_4 in a two-phase process. 28

3.2 Building of feature vectors. 30

3.3 Implementation scheme for the basic networks preparation in three steps. . 33

3.4 Implementation scheme for the simulations preparation in three steps. . . . 34

3.5 Simulation of adoptions. Either memes can be remembered with chances given by (a) the λ function from topical homophily or (b) equal chances. . 36

4.1 Distribution of the calculated similarity values. 39

4.2 Observed quantity of possible adoption events in blue and actual adoptions in yellow for each time value. 40

4.3 Quantity of users (a) and memes (b) related in the possible adoption events over time. 40

4.4 Distribution of the calculated topical fitnesses. 41

4.5 Distribution of average topical fitnesses of adopted memes (red/brown) and non-adopted memes (blue/cyan). 42

4.6 Distribution of the number of adoptions of each meme in the baseline. . . . 43

4.7 Adoptions counted to each meme over its lifespan. 45

4.8 Adoptions counted to each meme over its lifespan. 45

4.9 Average Kendall τ and their standard deviations obtained for topical homophily simulations in the upper blue line and uniformly random simulations in the lower red line, over the number α of memory slots. 47

4.10 Average meme lifespan precisions and their standard deviations obtained for topical homophily simulations in the upper purple line and uniformly random simulations in the lower orange line, over the number α of memory slots. 49

4.11 Average precisions and their standard deviations obtained for topical homophily simulations in the upper blue line and uniformly random simulations in the lower red line, over the number α of memory slots. 50

A.1 Hashtag co-ocurrence network degree distribution. 63
A.2 Mentions network in-degree distribution. 64
A.3 Mentions network out-degree distribution. 64

B.1 Number of ocurrences over average similarity, 100 bins. 65
B.2 Reproduced histogram of average similarities in the mentions network. Source:[13] 66
B.3 Distribution of average Jaccard indices over each user in a semi-log scale. . 66

List of Tables

4.1 The 10 most adopted memes in the baseline. 43
4.2 The top 10 longest meme lifespans in the baseline. 44
4.3 Kendall τ coefficient values obtained between the topical homophily simulated adoptions rank and the baseline adoptions rank. 46
4.4 Kendall τ coefficient values obtained between the uniformly random simulated adoptions rank and the baseline adoptions rank. 47
4.5 Summarized Kendall τ coefficient results between both types of simulations. 'Avg' stands for average, 'SD' for standard deviation, 'TH' discriminates the topical homophily simulations and 'UR' the uniformly random simulations, and the difference is taken between the average values. 47
4.6 Meme lifespan precision values obtained between the topical homophily simulated adoptions and the baseline adoptions. 48
4.7 Meme lifespan precision values obtained between the uniformly random simulated adoptions and the baseline adoptions. 48
4.8 Summarized meme lifespan precision results between both types of simulations. 'Avg' stands for average, 'SD' for standard deviation, 'TH' discriminates the topical homophily simulations and 'UR' the uniformly random simulations, and the difference is taken between the average values. 48
4.9 Precision values obtained between the topical homophily simulated adoptions and the baseline adoptions. 49
4.10 Precision values obtained between the uniformly random simulated adoptions and the baseline adoptions. 50
4.11 Summarized precision results between both types of simulations. 'Avg' stands for average, 'SD' for standard deviation, 'TH' discriminates the topical homophily simulations and 'UR' the uniformly random simulations, and the difference is taken between the average values. 50

A.1 Distribution of the users' frequency. 62
A.2 Distribution of the hashtags' frequency. 62
A.3 Top 5 hashtags in degree centrality. 62
A.4 Top 5 hashtags in eigenvector centrality. 63

Contents

1 Introduction **13**

1.1 Research Scenario 13

1.2 Problem Definition 14

1.3 Objective and Contributions 15

1.4 Methods and Challenges 15

1.5 Thesis Outline 16

2 Foundations and Related Work **17**

2.1 Social Spreading Phenomena 17

2.2 Topics of Information 23

2.3 Homophily 25

3 Topical Homophily on the Meme Spreading **27**

3.1 Model 27

3.1.1 Meme Adoption Process 27

3.1.2 Topics of Information 29

3.1.3 Topical Homophily 29

3.1.4 Topical Fitness 31

3.2 Data Handling 32

3.2.1 Tweets Collection 32

3.2.2 Basic Networks Preparation 32

3.2.3 Adoption Networks Preparation 34

3.3 Computational Simulation 35

3.4 Validation 36

3.5 Kendall rank correlation coefficient 36

3.5.1 Meme Lifespan Precision 37

3.5.2 Precision 37

4 Results **38**

4.1 Topical Homophily 38

4.2 Possible Adoption Events 39

4.3 Baseline Characterization 42

4.4 Simulations 46

4.4.1 Kendall Rank Correlation 46

4.4.2 Meme Lifespan Precision 48

4.4.3 Precision 49

5 Conclusions 52
5.1 Discussion 52
5.2 Contributions 53
5.3 Future Works 53

Bibliography 55

A Supplementary Material 61
A.1 Dataset Collection 61
A.2 Descriptive Statistics 62
A.3 Hashtag Co-ocurrence Network 62
A.4 Mention Network 63
A.5 Filtering 64

B Bias Discussion 65

Chapter 1

Introduction

Back in 1976, Richard Dawkins defined *meme* as a "unit of culture" [22], i.e., a small piece of information which could be an idea, belief, or even a pattern of behavior that is also analogous to a gene. This analogy means that it is subjected to evolutionary dynamics, such as reproduction, mutation and selection. Memes reproduce to transmit cultural information when people adopt ideas or behaviors as in a contagion pattern. Because it is thought to be hosted in each person's mind, it can be altered or incorporated with other ideas and hence mutate. As for the selection process, it is hard to reach a consensus about what happens to memes which survive or not. The cultural selection theory is the effort to grasp how the natural selection-based evolution principles promote culture change, but it spans through so many disciplines that epistemological and conceptual problems arise to the scientific point of view [21].

In the recent decades, a revolution took place in the core of the scientific method [34]. The so-called fourth research paradigm emerges from the conjuncture of massive data production – from a diversity of devices and simulators – and data access through networks and the web. It enables the scientific method to transit around theory generalization, computational reproduction and empirical validation in a fast pace, because of the range of new knowledge discovery possibilities opened by the available data. The remaining of the chapter is organized as follows: Section 1.1 introduces the main field of study of this work; Section 1.2 states our research problem; Section 1.3 presents our objective and contributions; Section 1.4 sums up our methods and challenges; Section 1.5 summarizes the organization of this document.

1.1 Research Scenario

A very important research area has been gaining attention in this period of data-intensive research. The network science [8] grounds itself in the graph theory in conjunction with the mechanical statistics to reveal, in a first moment, the structure and dynamics behind such an omnipresent formation as the networks. The concept of complex systems embraces systems formed by interacting parts with relatively simple behavior which show global patterns that defy reductionism. Such systems appear in many places in our world. This is also the case for complex networks [11]. Social networks commonly figure among the complex networks. This field is fundamental to the investigation of spreading processes in networks [50]. For instance, it defines two types of contagions: the simple contagion, which can succeed with only one exposition, common to biological mechanisms such as epidemic disease outbreaks and the complex contagion, which require reinforcement from multiple sources to trigger and are usually seen in social mechanisms such as the diffusion of innovations [16].

In the wake of the scientific progress of the complex systems, the computational social science was formed to tackle great modern challenges of our increasingly interconnected world, such as infectious diseases dissemination; economic and social crises; unethical use of communication technology; and other collective behavior phenomena [20] [40]. It also opened new opportunities by revisiting several concepts of the quantitative social sciences and promoting their integration with new models and techniques, also from the network science. For example, the triadic closure [33], a principle that participates in the formation of some network, or the cultural diffusion, a model created by Robert Axelrod [4] in 1997 that describes the interplay of social influence and homophily. Homophily is a long-standing concept in the social sciences, which reflects our tendency to interact with people that we see as resembling ourselves, i.e., as it was simply put in the seminal paper by McPherson et al. [45], "similarity breeds connection". Although interaction or connection has several meanings to this field, in the context of social networks, we are talking mostly of tie formation between individuals to represent whatever the social network proposed to abstract, e.g. friendship, work, recommendation or exchange.

1.2 Problem Definition

In 1992, Robin Dunbar found a correlation between the size of the neocortex and the social group size of primates [24]. This result points to a limitation of the cognitive functions, in the sense that the animal can not surpass a number of relationships due to the information overload. Gonçalves et al. [32] revisited this work in the Twitter social media platform. They concluded that indeed the users cannot hold more than a certain number of stable social relationships – those which they use to communicate actively – in an online social network. Michael Goldhaber anticipated [31] that we live in an attention economy, meaning that attention is the currency that people base upon to decide how to navigate through the information ambiance.

The online social network's users receive much more information that they can manage to interact with. An important work by Weng et al. [60] found that the memory of these users is a strong constraint in this system. Seeking elements to explain the long-term persistence of very few memes, they observed that, while the proportion of meme diversity in posts over the network grows, the messages of each user cannot cope with this growth. Therefore, the diversity of these messages would remain practically constant in comparison. This result means that users have limitations to deal with the overload of information. It indicates a selective pressure among memes, in the sense that users have to choose which memes adopt and share, due to their limited cognitive capacity. In the present research, we attempt to understand how memes are selected in a restrained attention environment, such as the online social networks, by asking ourselves how can this memory limit be embodied into a model. In this context, memes can be phrases, images, videos or hashtags. Our work is limited to hashtags.

In this setting, we pose the following research question: *how the homophily with users interplays with the memes competition for attention?*

Since there is an intensification in the usage of online social networks in contemporary communication, this research challenge is increasingly imperative. Firstly, we face a growing demand for curating the instruction and claims in these platforms [41]. The scientific comprehension of the underlying diffusion processes should prove itself a paramount contribution to policy-making, regulation and ultimately the integrity of institutions since this new technology is affecting individual decisions to an unknown degree so far [59]. For instance, it could lead to the development of methods to identify effects in opinion propagation, such as the echo-chambers [6]. There are also possibilities to accelerate the broadcasting of public-interest information, such as missing-person searches, health alerts or disaster relief, as well as industry-relevant implications, such as the engineering and control of mechanisms for viral marketing.

1.3 Objective and Contributions

Our objective is to model a function for the fitness of shared memes – concerning their reuse in messages produced in online social networks – in correspondence with data obtained empirically.

To advance towards it, we have investigated the hypothesis that homophily plays a role in a meme fitness measure. This investigation was conducted with a novel model which captures a simplified memory behavior from the online social network users. It also produced empirical indications from a real dataset that the competition exists and that a specific formulation of homophily has influence in this scenario.

We expect that our model will contribute, in the immediate scope, to applied populational studies for prediction and meme control and, to a mid-term scope, supporting the mathematical formalization of meme evolution in the social spreading phenomena.

1.4 Methods and Challenges

This research endeavor comprises an abstraction of the meme adoption process to accommodate a fitness function which characterizes how likely memes are remembered. This abstraction has been based on aspects of different models of the literature, such as the two-phase adoption from Gleeson et al. [29] and the discrete memory slots from Qiu et al. [51]. While these works account for careful representations of memes spreading in a highly heterogeneous fashion, as observed in the online social networks, they do not define a direct competition among memes to obtain attention from the users, in the sense of one meme having characteristics which are more attractive to a specific user.

We modelled this advantage in the competition for attention as a particular case of homophily. The topics of information are clusters of memes, which allow a quantitative characterization of the homophily, called topical homophily. Homophily is observed when users interact more with their similar peers - which, in our case, means users interacting more with those that use more memes from the same topics. We have built a fitness function based on this kind of meme adoption. Our experiments with this function are based on simplifications of time, which can potentially lose parts of the involved dynamics, as memes are ephemeral entities and their spreading processes can take place in very short periods.

We performed a study of meme adoption with the proposed fitness function using real data, Meme spreading simulations are difficult because a tiny number of them become massively popular, while most die out quickly, what characterizes an erratic dynamical system.

1.5 Thesis Outline

This thesis is organized into six chapters. Chapter 2 inspects the foundations of our research and related work. Chapter 3 details the models, definitions and methods that we used to reach our research question. We provide a simplified model to be developed later in a computational experiment, with a real dataset that will help us to understand how topical homophily exerts influence in meme adoption. Chapter 4 details our results. Chapter 5 evaluates the results and methodology, recapitulates our main findings and identifies the opportunities for future works.

Chapter 2

Foundations and Related Work

The present research comprehends a social network analysis and this task involves investigating the relations between the information that individuals in a network share and the structure of connections among them. We start in Section 2.1 by examining how social influence investigations characterized the social spreading phenomena and subsequently layed the groundwork for various mathematical formulations to the spreading processes. Section 2.2 addresses the conceptual basis of content classification, in terms of topics of information, or clusters of semantically related pieces of information. Finally, in Section 2.3 we discuss the concept of homophily, how it is used in the context of social dynamics and the later developments.

2.1 Social Spreading Phenomena

The social spreading phenomena comprise the diffusion of information and its inherent dynamics over social networks. Online social media, such as Facebook[1] and Twitter[2], offer large-scale sources of data concerning human behavior and relations [38], which enabled robust lines of investigation in the social context. In these platforms, the social network is defined mostly by users linking to each other through *following* or *friendship* relations, and the information analyzed are mainly the messages exchanged by these agents in the network.

We remind that the concept of the meme was defined in an evolutionary perspective, so that it is subject to reproduction, selection and mutation [1]. In the context of the research of online social networks, memes are treated as transmissible units of information [26] and they are embedded in the messages posted by each person in the network, e.g., small phrases, images, videos, audios, URLs or hashtags. We are interested in hashtags due to their capacity to tag posts with one or few words, making them easier to track than a short phrase [43] and harder to mutate [57], which allow a clearer handling of semantic association.

[1] https://facebook.com

[2] https://twitter.com

One of the most studied online social networks is Twitter [65]. The platform makes publicly available about 1% [47] of all the posted messages through its application programming interface. Its messages are called *tweets*. It is known as a microblogging service because there is a size limit to each tweet: it rose from 140 characters to 280 recently, in September of 2017. This is an interesting feature in comparison to blogging platforms, as the bounded size eases the automated analysis of the content in each message. It also has a formal mechanism to reshare the messages named *retweet*, although it is possible to reshare a fragment of a tweet without using this mechanism. Twitter's social network is established with the *following* relationship so that the follower is notified about the messages posted by the followee. The directionality of this network is also an appealing attribute to conduct scientific research, as it is easier to generalize models from directed networks to undirected networks than the contrary. These characteristics support a rapidly changing environment of messages, which makes Twitter a good source of data to examine the social spreading phenomena.

The online social networks bear intricate structural patterns, such as power-law degree distributions [46]. It means that there are very few hub nodes which are much more connected than the others, while the majority of nodes display a tiny number of connections. The hubs retain a broader audience than the majority of other nodes, as the messages they emit are received by each of the numerous users around them. Therefore, they amass much more influence than other people, as largely spread messages will likely pass or even be started by them. A reasonable deal of the effort in the field of computational social science revolves around the problem of identifying key influencers in a social network.

This problem, closely related to our investigation, is known as the influence maximization problem. It was first posed in the algorhtimic sense by Domingos et al. [52] in 2002. Kempe et al. [36] achieved a noteworthy advance in the theoretical perspective by showing an efficient approximation on how to maximize the number of activated nodes given an initial number of activations. They also demonstrate the equivalence of two diffusion models for the diffusion of ideas, the threshold model and the cascade model. Even though the problem had such algorithmic development, many empirical endeavors were conducted to recognize the characteristics of heavy influencers in realistic settings [42] [3] [17] [5] [53].

One of the most intriguing aspects yielded by the unbalanced influence of a social network is its popularity distribution. Notably, most of the memes stop to be reshared quickly, while a few of them outlive the rest by being continuously spread through the network. A specific spreading pattern occurs when edges are successively activated in a cascade-fashion process [30]. We say these memes become viral, due to the similarity to a virus diffusion in epidemic models. We use Figure 2.1 from [30] to depict the difference between viral diffusions and broadcasts, as both of them can be associated with popular memes. From the empirical point of view, the modelling of the information cascades proved to be rather intricate, as seen in [62], [7] and [18]. The cascade sizes vary greatly, as does the activity patterns of the users, and these cascades are also very sensitive to the network structure.

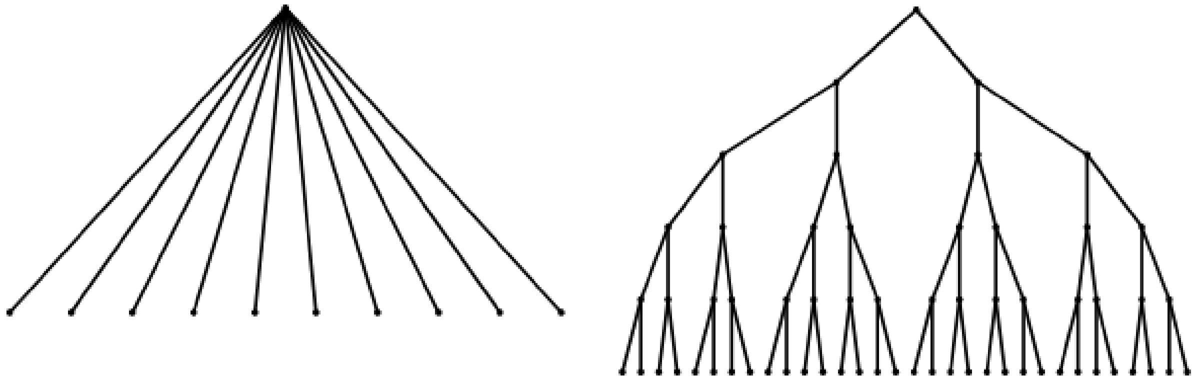


Figure 2.1: Difference between broadcast (left) and viral diffusion (right). Edges represent meme adoptions. Source: [30].

Zhang et al. provide an extensive review [64] of several endeavors to approach the social spreading phenomena, from pragmatic to theoretical viewpoints, ranging through various social media platforms and different modelling choices. We use this review as a reference to what kind of modelling direction we choose to follow.

For the sake of investigating if memes possess innate traits that make some more attractive than others, we have opted to guide our investigation by the cascading models and equivalent forms, because the cascade size - resulting from consecutive adoptions of a meme - is a good indicator of how well it fares in this dynamic landscape. This class of models is built to describe the sequential activation of nodes in a network, and thus it often features processes with a strong time-dependence assumption. A single change in the middle of the process may deeply alter the cascade structure and size. This is a defining choice in the efforts to tackle the modelling challenge in the problem of information diffusion, because as seen in [64], the problem may be approached by different kinds of models. For instance, models with the Markovian property are dubbed as memoryless because they generally assume the next state of the system only depends on the previous.

The experimental findings of Weng et al. led them to create an agent-based model, assuming a static social network, which considers users with a time-ordered list of posts, each containing memes. A single meme may appear in more than one post of the same user. Users are notified only about the memes posted of other users they follow. They decided to treat the memory of each user as discrete slots being constantly updated, so each person, exposed to new memes, must select which memes will be kept in mind, between old and new ones. This mechanism is illustrated in Figure 2.2.

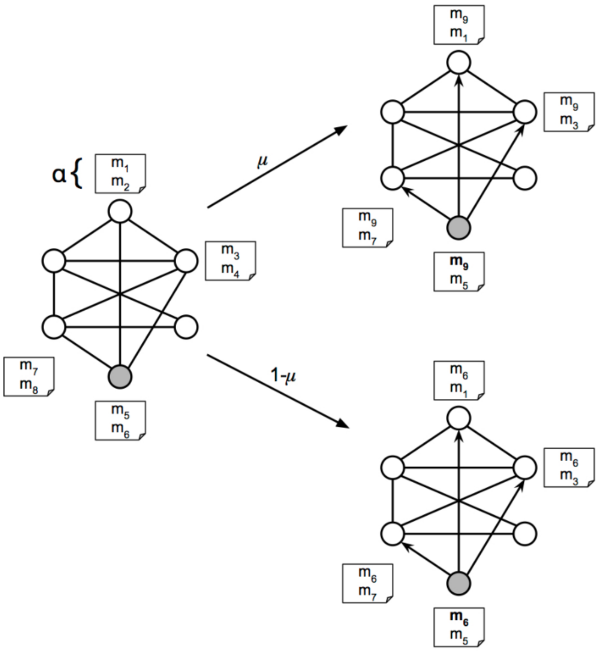


Figure 2.2: Each user has two discrete memory slots. At a given time, the shaded user may create a new meme (m_9) with probability μ or reshare an old one in his memory (m_6) with probability $1-\mu$, and his followers may adopt it by replacing one of the memes in their memories. Source: [51].

A competition among memes is established and those which succeed outlive the forgotten contenders. Through this model, the memory slot mechanism and the social network structure are combined to approach the real-world heterogeneous popularity. The authors analyzed the model against a dataset collected in four months, amounting up to 120 million retweets, 12.5 million users and 1.3 million hashtags, in order to validate the outputs of the simulated model. Even though achieving noteworthy results in the explanation of the heterogeneity in popularity and persistence of memes, the model still cannot clarify the conditions that trigger a phase transition in which a single meme diverts the node activations towards itself, producing a large cascade of retweets.

Pursuing a mathematical formalism to the emergence of cascades in a memory-constrained setting, Gleeson et al. applied branching processes to describe the criticality of popularity distributions [28]. The critical point is a physical concept from thermodynamics which marks the condition for a phase transistion for a system in equilibrium to disequilibrium. In the context of social networks, this equilibrium is a state of meme sharing which sequencies of adoptions are modest and no meme becomes viral, as opposed to the disequilibrium represented by a huge cascade of adoptions. A branching process is a random process that proceeds through generations, each of which has some number of individuals that produce more individuals in the following generation, according to some distribution. In the study of Gleeson et al., the individuals were mapped to edge activations, the equivalent to meme adoption, in the social network. This work presumes a setting with no innovation, i.e., the memes being shared already exist in the network and users are not allowed to create new memes.

Under this hypothesis, their equations can produce the same power law distributions of popularity encountered in previous empirical investigations, e.g., [7]. When there is no original meme during the evolution of the system, the exhibited critical behavior is provoked by the competition for the users' limited memory. Despite the meaningful insights given by this analytically tractable theoretical framework, the model considers only simple extensions to the update of memory slots. It was not much concerned on how memes are accepted or rejected into memory, and if any meme could have an advantage in the competition.

Gleeson et al. proceeded to generalize the equations of competition-induced criticality in the next endeavor [29]. The mapping of a branching process to the cascade model is developed to approximate several distributions in the dynamics. Their model regards a meme adoption process in two phases. Firstly, memes deemed interesting are taken by the user into his memory, from the possible memes at that time - those being shared by the user's neighbours. In the second phase, the user will select one of the memes in his memory or create a new one and share it with the social network.

In the first phase, memes are remembered with an equal probability λ and the limitation of memory is explained with a distinct structure. Instead of a fixed number of memes being remembered at each time step, each meme has an expiration time. This time is drawn randomly under a memory-time distribution, a probability density defined to control the duration of memes in memory. The method is convenient to run simulations and indirectly estimate meme duration in memory in a macroscopic observation of a real dataset, especially because it is not feasible to measure exact amounts of time that each person remembers a meme.

Furthermore in the second step - the election of memes to be posted in a message - users may generate original memes, instead of picking one in their memory, with a chance given by the innovation rate μ, an important parameter in this system. The decision of which meme is shared is also based on the frequency that the user sends his messages. The rate in which users produce messages is heterogeneous, so the authors studied how this rate affects the process by classifying it accordingly to the out-degree of the users. That proportionality makes sense because users with a bigger audience are likely to take advantage of the bigger influence exerted with the increase of their own messaging effort.

They performed numerical simulations to examine the quality of the results obtained to the equations of cascade sizes, i.e., the population of memes and the prediction performance in two network topologies. The simulations reasonably fit the curves foretold by the asymptotic analysis for large values of the age and popularity of memes. The age of a meme is the measurement of how many days it lasted. In addition, the model is confronted against real-world data, gathered for a year, with careful adjusts to the considerations made in theory. This confrontation results in the correct prescription of distributions given by the model to what was seen in the empirical settings.

Nevertheless, this robust work aims to be a neutral model, in the sense that all memes are deemed interesting with constant probability given by the parameter λ. The null models, models that match structural properties of interest, are fundamental to set a baseline which complement data-driven approaches, and their work promotes this fact. Still, there is an opening challenge we intend to work in this thesis, which is the study of how λ varies to each meme and each user, the so-called fitness of memes, or their inherent advantage in the competition to be retained in one's memory.

A study seeking to discern meme features behind their success was done by Qiu et al. [51]. The authors proposed a property named *quality*, intrinsic to each meme and proportional to the likelihood of adopting it from other people. It does not represent an advantage in the dispute for attention, but weights to how probable the meme will be posted when it has already been taken into the user's memory. To tie this memory adoption aspect, the authors also consider the system has *discriminative power*, a correlation between quality and popularity. A high discriminative power means that the social network users are able to identify memes of high quality. From those assumptions, they set out to understand how the ability of people to distinguish quality in information may affect its virality.

An agent-based model was designed to reproduce the meme adoption mechanism, rendered in Figure 2.2. Each agent has a number of memory slots, constant to all of them and given as a parameter to the system. These memory slots are updated in chronological order, i.e., whenever a user is exposed to a new meme, if his slots are all occupied, the oldest meme will be forgotten and replaced by the new one. A randomly chosen agent, at each time step, may generate a new meme (with chance μ) or take one from his memory accordingly to its quality to generate a new message. Every created meme have its quality randomly drawn in the uniform interval between 0 and 1. Then, simulations were performed through synthetic networks with variations in the number of memory slots and the information load - the average number of memes being exposed to a user.

The model calibration with empirical inputs was carried out with several data sources. The first was the Twitter Streaming API, which provided hashtag popularities and a practical measure of the innovation rate, as the fraction of tweets over the sum of tweets and retweets. Then, the microblogging service called Tumblr was used to estimate the distribution of the number of memory slots, by recording how many stops were made when people scrolled batches of at least 500 pixels during sessions in mobiles. These stops act like a proxy for an attention effort, supposing that people would not pay attention to content when they are rolling in the screen. Finally, the researchers collected data from a rumor tracking project named Emergent, which is deactivated now. This project was created to check the integrity of claims from articles shared on Facebook. Trustworthy claims were classified as high-quality articles and poor or false claims as low-quality. This binary designation was fed to the model as an empirical measurement of quality.

These works investigating and mathematically structuring the spreading processes carve a course of valuable perception on the effects of the local memory for the emergence of collective patterns of behavior. We should pay attention to them in order to ascertain assessments of the meme dynamics.

2.2 Topics of Information

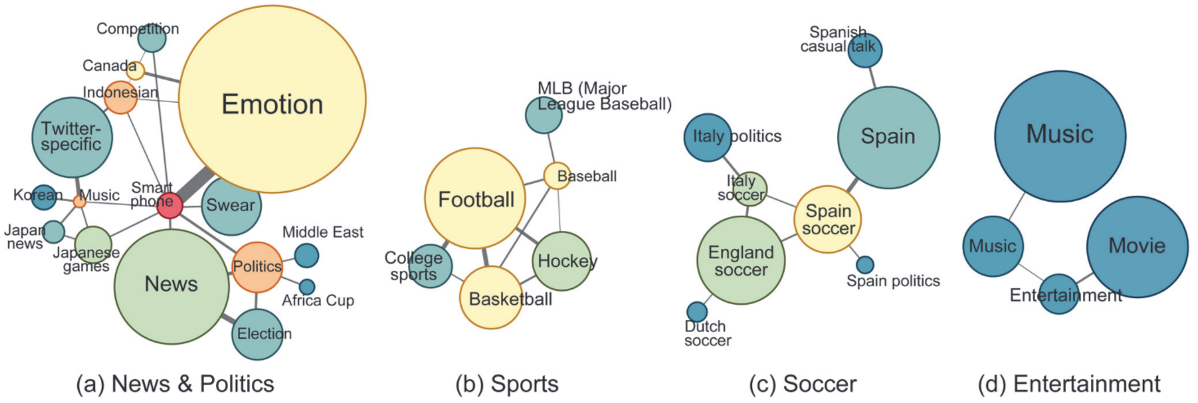
When studying the flow of information, it is possible to focus on the information itself, rather than the propagator agents. Social network users may want to participate in trending movements, redirecting their attention to emerging groups of memes [35]. This means the evolution of messages' content takes part in the behavior of the agents. It is important to consider how to characterize the information flowing and changing in the networks and the interaction among their distinct kinds. We consider a manner of characterizing this content with topics of information, clusters of semantically related pieces of information, which in our work are hashtags.

A study deals with the interaction of different topics in conjunction with the limited attention scenario [19]. Ciampaglia et. al evaluated the demand and supply of information across different topics, to support the vision of an attention economy in this system. By examining attention bursts in Wikipedia traffic, they confirmed that the emergence of high traffic in a specific topic - i.e., demand - precedes the creation of content related to that topic, which is supply in this context. The collective attention shifts reinforce the effectiveness of trending topics changing local behaviors.

In Twitter, the overwhelming information generated on a daily basis is often easier to group than to isolate in different kinds. Earlier attempts to arrange together different pieces of information took advantage of the semantic similarity among memes. Sayyadi et al. [55] isolated keywords from documents to build a graph of co-occurrence, which have low text frequencies filtered out. Then, they performed a community detection to group the keywords. Cataldi et al. [14] designed a method to spot emerging topics in real time using keywords with high text frequency and usage in defined time intervals. They also built a topic graph representing the semantical relationships between co-occurrent keywords which generated the topics. These are good directions to group keywords, but these keywords were not necessarily hashtags, terms created expressly to identify the content of messages.

Ferrara et al. [26] tackled the challenge of defining groups of memes and assigning memes to groups. They developed a framework to cluster memes by several similarity measures considering four different types: hashtags, URLs, mentions and phrases. They defined four base similarity measures and studied combinations of them to determine the best performance for the comparison between hierarchical clustering and K-means clustering techniques. With a careful experimental setup and cross-validation, they found out that the hierarchical clustering - a method often used to detect community structures in networks - returns a higher trade-off between the number of clusters and their quality.

By improving the approaches to cluster memes or keywords, Weng et al. [61] produced a work about the relationship between social influence and topics of information. They considered hashtags facilitators for information retrieval, which can connect people with shared interests in a space with a high diversity of content. After the application of a noise filter to hashtags, a network of hashtags was built by linking pairs that figure together in at least three messages, resulting in a **hashtag co-occurrence network**. By also assuming that semantically close hashtags may appear together more likely than those semantically distant, the resulting network is densely connected and counts with community structures. They performed a community detection task in this network with the Louvain method and assigned a topic of information to each cluster of hashtags. Some of these clusters are illustrated in Figure 2.3.



(a) News & Politics (b) Sports (c) Soccer (d) Entertainment

Figure 2.3: Examples of different topic clusters and how they are connected after being retrieved from the hashtag co-occurrence network. The size of each circumference indicates how many hashtags are inside the cluster. Source: [61].

Weng et al. also designate the interests of each user as a function of the topics containing hashtags used by that user. This is an appealing representation because it makes possible to understand how the diversity of themes affect the communication and social influence exerted by users. It can also be used to analyze the content diversity in hashtags since a hashtag in several topics is closer to more diverse content. There are also two experimental setups to predict hashtag popularity and social influence metrics. Finally, their findings point to messages more diversified hoarding more popularity, as well as users focused on fewer topics.

Another study of influence which identifies social network users by the topics they are talking about was done by Bogdanov et al. [12]. They focused on evolutionary aspects of the topical behavior of users and how these aspects can be used to predict influence in social media, even improving results obtained from structural network metrics, such as centralities. It is another indication of the value in characterizing interest of the users with topics of information.

After inspecting how the topics of information are formed and how they can be used to distinguish content interests in social networks communication, we will now examine some fundamental notions behind our hypothesis and a very important work that interlaces topical interests and homophily.

2.3 Homophily

The tendency of people of linking themselves to those they judge similar, also known as homophily, has been studied for several decades in the social sciences [56]. The notion of homophily has been generalized from the older idea of triadic closure [37]. Considering three nodes in a graph, this concept can be seen as the impulse to form a third edge in a triad of nodes possessing already two edges, e.g., how likely one forms friendship with the friend of a friend. Even though this may not be ubiquitous in complex networks of large-scale [25], it is a simplification used to advance the understanding of social networks formation for many years.

The social media came to offer not only enough data to statistically deepen our discernment of the predominance of these ideas in social systems but also several other options of ties formation among people [23]. Online social network users form ties amongst themselves considering static socioeconomic or geographic characteristics, but also traits regarding their activities or the content they like to consume and communicate over.

In particular, Weng et al. [63] looked at how much the creation of links can deviate from a triadic closure when people decide to connect after reading messages that were spread in the network. They found that popular users have the capacity of shortening the paths traveled by messages due to the generation of intense traffic around them. These users draw links to increase the efficiency of the information diffusion, so that the network has an evolution setting contrary to the intuition of local growing shaped by triadic closures.

In fact, the ability of using social influence to create interaction is another principle closely related to homophily. The Axelrod model for cultural diffusion [15] poses a mechanism in which both homophily and social influence reinforce each other enabling, counterintuitively, a consistent diversity of the cultural configuration in a group of people.

Even though this is another abstraction hard to find correspondence over significantly large realistic scenarios, it sets the foundations for important investigations such as the one done by Aral et al. [2]. They dissect the predominance between social influence and homophily by exhibiting a method to separate when a complex contagion is an effect of one or the other. An experiment is also executed in a network of more than 27 million users of a messaging platform, based on their adoption of a mobile product. The results have shown that the number of adoptions triggered by homophily can be severely misjudged by methods based on the network structure, which attribute them to a social influence instead.

We are interested in understanding how homophily affects the adoption of memes, so it is essential to characterize it with behavior traits rather than structural properties of the social network, like centrality indexes. In this scenario, the topics of information open a compelling opportunity to work with. Cardoso et al. [13] supply a framework to describe homophily based on topical interests. As the topics of information groups hashtags, this framework allows a quantification of how similar the social network users are based on how they use hashtags from the same groups. This is a specific representation of homophily they call **topical homophily**. The topics are produced similarly to the model of Weng and Menczer [61] described in the previous subsection. They analyzed several messages in Twitter and concluded that higher topical similarity yields stronger interactions between users, on average, and that this measure even allows link prediction to some extent. We consider the framework they used as a simple yet powerful way to establish homophily in the online social network dynamics.

In the next section, we will describe how we can take advantage of this framework in the meme adoption context.

Chapter 3

Topical Homophily on the Meme Spreading

In this chapter, we recall the important aspects observed in the literature about the problem; state our hypotheses; and present our proposal. This proposal to obtain empirical evidence of the influence of topical homophily in meme adoption is built from a simplified abstraction presented in Section 3.1. Section 3.2 presents our process of gathering data from Twitter and preparing it to our analysis. Finally, we present the design of a computational experiment in Section 3.3 and how we validate the model through this experiment in Section 3.4.

3.1 Model

In this Section, we describe how our model is thought. Consider a directed social network where the users establish a relationship of *following*. This kind of relationship works like a passive subscription, in which each user receives messages from those he follows. People being followed can be called *followees* or neighbours. Each message contains memes which, in our case, will be restricted to hashtags. When producing a message, the user may generate an original meme, or adopt one of the previously seen and embed it in the message. The later is the case that concerns this research.

3.1.1 Meme Adoption Process

The meme adoption process is thought to happen in two phases, based in Gleeson et al. [29]. Let us assume we have the time split into a discrete set. Suppose that a user u posts a message in time t after receiving all messages posted by the other users he follows at the previous time $t-1$. He can adopt memes from previous messages that he remembers, but not all of them are remembered because u has a memory limit. At the first phase, some memes are stored in his memory, while some of the ones already in memory may persist or not. In the second phase, u composes a message by inserting one or more memes from his memory or generating new memes, which are not found in the messaeges of his followees. We are interested in the former case. If a meme h posted at time $t-1$ by the neighbours of u is successfully stored in his memory and then used in a message composed by u at time t, we say that u **adopted** the meme h.

While we can observe when a meme h is posted by u right after it has been posted by his neighbours, it is not feasible to gather data about which memes are remembered by him in such a dynamical context as the online social networks. So we take a simplified memory model, inspired by the model of Qiu et al. [51], to simulate what happens in the first phase of meme adoption. We consider that each user has a number α of slots that can be used to store memes. At each adoption event, i.e., when the user receives possible memes to adopt from his followees, each meme may be accepted into his memory with probability λ. In the null model provided by Gleeson et al. [29], the parameter λ is considered constant, which means that no meme holds an advantage in the competition for the attention of the user, a limited resource.

Our hypothesis is that memes have different probabilities to be remembered according to a fitness function. This work is focused in studying what is behind the *fitness* of a meme and how it stands in the competition. Therefore, our model defines a function λ to weight the adoption probability among memes. Such probability is distributed among memes eligible for adoption from a specific period of time. We summarize how we model the adoption process in Figure 3.1.

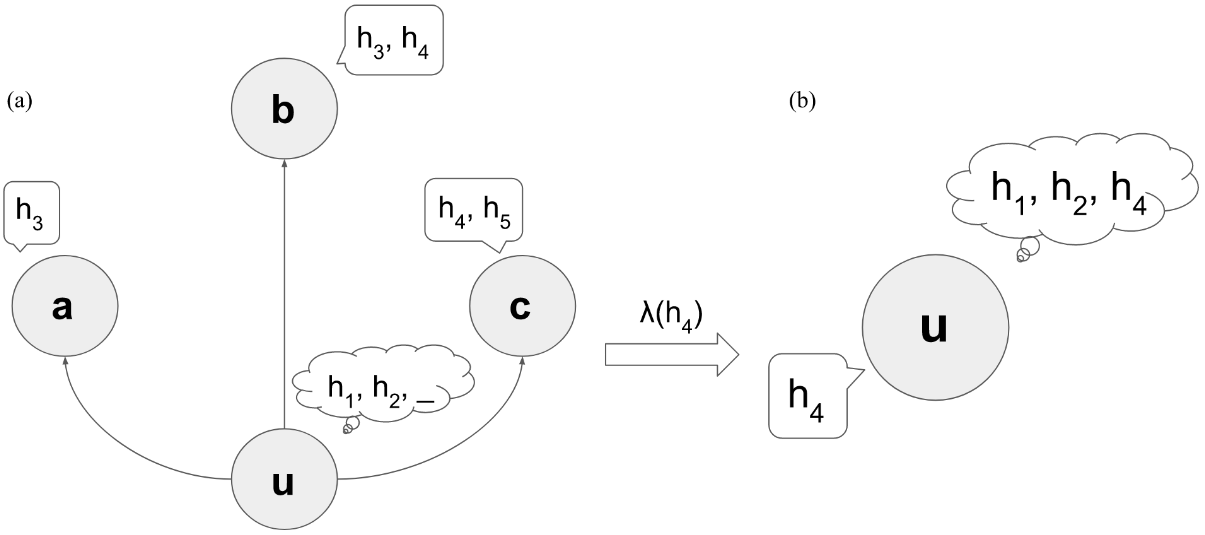


Figure 3.1: Adoption of the meme h_4 in a two-phase process.

Let us take an instance of an online social network where user u follows users a, b and c, he has $\alpha = 3$ memory slots and adopts meme h_4 in a two-phase process. In the first phase (a), user u incorporates one meme between h_3, h_4 and h_5 from his followees to one of his three memory slots. h_4 is drawn with probability $\lambda(h_4)$, and in the second phase (b), user u chooses to use it in his message from the three possible memes.

Our investigation wants to verify whether the topical homophily affects the process, considering hashtags as our selected meme type to the object of study. We further characterize how the topical homophily can constitute such a function λ denoting how likely a meme is remembered.

3.1.2 Topics of Information

As seen in Bogdanov et al. [12], it is possible to identify users in terms of their topic interests, or the topics of information which contain their memes. The topics, in turn, can be detected from communities in a hashtag co-occurrence network. This is an undirected network obtained by linking nodes representing hashtags that appear together in any given message. Its edges may also be weighted by the number of joint occurrences of each meme. In the example given in Figure 3.1, if h_3 and h_4 appeared together in three messages, their corresponding edge would have weight equal to 3.

We then perform a community detection in this undirected weighted network with statistical inference and modularity maximization techniques [49]. Each module returned by this process is designated as a topic of information, in a procedure defined by Weng et al. in [61]. If C(h) are the communities which contain h, then the set of topics of information of h is given by C(h).

That is, a module or community in the hashtag co-ocurrence network is a topic of information, so that each hashtag is attached to the topics accordingly to all the communities containing it. We opted to not use the singletons, i.e., the nodes isolated from communities. Once this is done, we can use the topics to identify the interests of users.

3.1.3 Topical Homophily

We consider a discrete set of time $T = \{1, 2, ..., n\}$ as our observation window. At each time t, a **possible adoption event** occurs for user u whenever a hashtag was used by a neighbour of u in time $t-1$. We define in our notation a set $N(t)$ of all neighbours which caused possible adoption events in time t.

At this point, we also reproduce the work of Cardoso et al. [13] to make use of the topical homophily. We chose their quantitative framework due to its simplicity and the versatility to deal with very different memes, but opposed to their general measurements, we adapt it to the specific situations of meme adoption.

In the same fashion of their model, each user may have his set of topics of interest expressed by a feature vector. But we build different feature vectors for u and his neighbours. To u, we acumulate all hashtags posted from 1 to $t-1$ to identify his interests, and to his neighbours, only those of $t-1$.

Let us name T the set of detected topics and $H(u)$ is a multiset composed by the hashtags used by user u, so that each hashtag has a multiplicity to reflect its usage. The topical interests of u are a feature vector $f(u)$ given by

$$f_i(u) = \sum_{h \in H(u) \cap T_i} \frac{\text{multiplicity of h in } H(u)}{\text{\#topics with h}} \tag{3.1}$$

where each position i corresponds to a detected topic $T_i \in T$. It means we weight the number of uses of each hashtag h over the number of topics which h pertains.

If we build a feature vector for each neighbour $v \in N(t)$ but using a multiset $H(v)$ with hashtags used during time $t-1$, it is possible to calculate the similarity $sim(u, v)$ between a user u and v, based on their topical interests, as

$$sim(u, v) = \frac{f(u) \cdot f(v)}{\|f(u)\| \|f(v)\|} \tag{3.2}$$

that is, the cosine similarity between their feature vectors. It is possible that there are null feature vectors if all the hashtags used are singletons. If we set this similarity to 0 when there is a null feature vector, it becomes well defined to ponder how close is the topical interests of neighbours. This is a way to position users regarding the content in their messages proposed by Cardoso et al. [13]. But as we apply it to adoption situations, by considering different sets of hashtags between the agent of adoption and his neighbours, we want to express that the agent has coherent topic interests and do not have to account for past interactions with his neighbours.

We will, again, build this concept by example. Consider we have two other past hashtags h_1 and h_2 and the five hashtags are grouped in two topics, given by $T_1 = \{h_1, h_3, h_4\}$ and $T_2 = \{h_2, h_4, h_5\}$. This means that our feature vectors will have two positions. Let us suppose that, just like it is rendered in Figure 3.1, we are in a time t such that we want to know the similarity between u and his three neighbors that used hashtags in $t-1$, so $N(t) = \{a, b, c\}$.

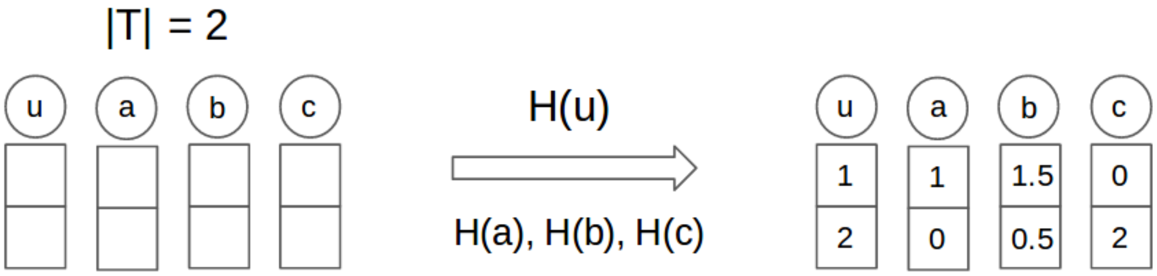


Figure 3.2: Building of feature vectors.

Let us say that, in our entire observation from time 1 to $t-1$, u used hashtags such that $H(u) = \{h_1, h_2, h_2\}$. If we consider the neighbours used the hashtags as showed before in Figure 3.1, $H(a) = \{h_3\}$, $H(b) = \{h_3, h_4\}$ and $H(c) = \{h_4, h_5\}$, we build feature vectors by applying Expression 3.1 as:

$$\begin{cases} f(u) = (\frac{1}{1}, \frac{2}{1}) = (1,2) \\ f(a) = (\frac{1}{1}, 0) = (1,0) \\ f(b) = (\frac{1}{1} + \frac{1}{2}, \frac{1}{2}) = (1.5, 0.5) \\ f(c) = (0, \frac{1}{1} + \frac{1}{1}) = (0,2) \end{cases} \tag{3.3}$$

We summarise the steps of initializing feature vectors with positions equal to the number of topics, defining the multisets of hashtags and filling each position based on hashtag pertinence in topics belongs to which topic with a illustration in Figure 3.2.

In our example, we calculate the cosine similarity yielded by u and his neighbours with approximations as defined in Expression 3.2:

$$\begin{cases} sim(u,a) = \frac{(1*1+2*0)}{(2.23)(1)} = \frac{1}{2.23} \approx 0.45 \\ sim(u,b) = \frac{(1*1.5+2*0.5)}{(2.23)(1.58)} \approx \frac{2}{3.52} \approx 0.57 \\ sim(u,c) = \frac{(1*0+2*2)}{(2.23)(2)} = \frac{4}{4.46} \approx 0.90 \end{cases} \tag{3.4}$$

3.1.4 Topical Fitness

Based on the found similarities, we define the topical fitness of a meme h to a user u as

$$TF_u(h) = \frac{\sum\limits_{v \in N_h(t)} sim(u,v)}{\sum\limits_{v \in N(t)} sim(u,v)} \tag{3.5}$$

where N is the set of neighbours followed by u. Back to the example of Figure 3.1, $N(t) = \{a, b, c\}$, and $N_h(t)$ those which posted the hashtag h (e.g. $N_{h_4}(t) = \{b, c\}$). Therefore, this is the fraction of similarity between u and neighbours who have used the meme h. We will study this measure of topical fitness by assuming the adoption probability is proportional to it, i.e., $\lambda(h) \propto TF_u(h)$. This is possible by applying a normalizing factor of the topical fitness considering all memes related in the possible adoption events of t, so that it sums to 1.

To further clarify it, we will move back to the example based in Figure 3.1. We have already determined the topics of information and calculated the feature vectors between u and his followees, taking into account memes posted by them in the desired time instant. The similarities yielded are $sim(u,a) = 0.45$, $sim(u,b) = 0.57$ and $sim(u,c) = 0.90$, so when analyzing TF_u over the set of memes $\{h_3, h_4, h_5\}$, we would have

$$\begin{cases} TF_u(h_3) = \frac{0.45+0.57}{0.45+0.57+0.9} = \frac{1.02}{1.92} \approx 0.53 \\ TF_u(h_4) = \frac{0.57+0.9}{0.45+0.57+0.9} = \frac{1.47}{1.92} \approx 0.77 \\ TF_u(h_5) = \frac{0.9}{0.45+0.57+0.9} = \frac{0.9}{1.92} \approx 0.47 \end{cases} \tag{3.6}$$

To shape these numbers to a probability distribution, we normalize them by multiplying by the inverse of the sum of these three Topical Fitnesses. Then we would remain with $\lambda(h_3) = \frac{0.53}{1.77} \approx 0.30$, $\lambda(h_4) = \frac{0.77}{1.77} \approx 0.43$ and $\lambda(h_5) = \frac{0.47}{1.77} \approx 0.27$.

This is how we establish the probability of adopting a meme based on the topical homophily when an adoption can happen. Moving forward, we planned a computational experiment over a real dataset to investigate the behavior of the proposed function λ. Its outline is detailed in the next sections.

3.2 Data Handling

In this Section, we explain how the data is gathered, pre-processed and prepared to our computational experiment.

3.2.1 Tweets Collection

To properly investigate the effects of topical homophily we have retrieved real data from Twitter. Its Streaming API makes about 1% of its tweets available [47]. To track the meme adoption we exploited two mechanisms. The platform includes a formal mechanism of adoption called *retweet*. It is also possible to quote the entire message without this mechanism [9], i.e., people can refer to pieces of messages with their own messages without the explicit retweet label. Both cases were treated equally as adoptions.

The social network, behind the analyzed messages, is inferred from the mentioning relationship instead of following. A user mentions another user in his message to call attention to that message, e.g., when one message replies another. Previous works [13] show that the network formed by this kind of relationship yields similar significance of interaction properties regarding homophily.

Tweets were collected with the Streaming API between September 18th and October 30th of 2017, totalizing a 45-day observation window. We chose the Portuguese language to filter them and used 137 keywords concerning brazilian politics, in order to narrow down a subject. We chose this subject since we judged that people are usually highly participative in it and also take positions. Nevertheless, hashtags about various other subjects and several other languages were retrieved from the entities annotated in each tweet, and we did not considered alternative memes embedded in them, as figures or text inside images. For instance, many of them are in the Spanish language and talk about the situation in the Catalunya province, which was a recurrent theme of debate back at that time.

3.2.2 Basic Networks Preparation

Figure 3.3 summarizes the preparation process of the dataset to identify the topics of information and to produce the mentions network.

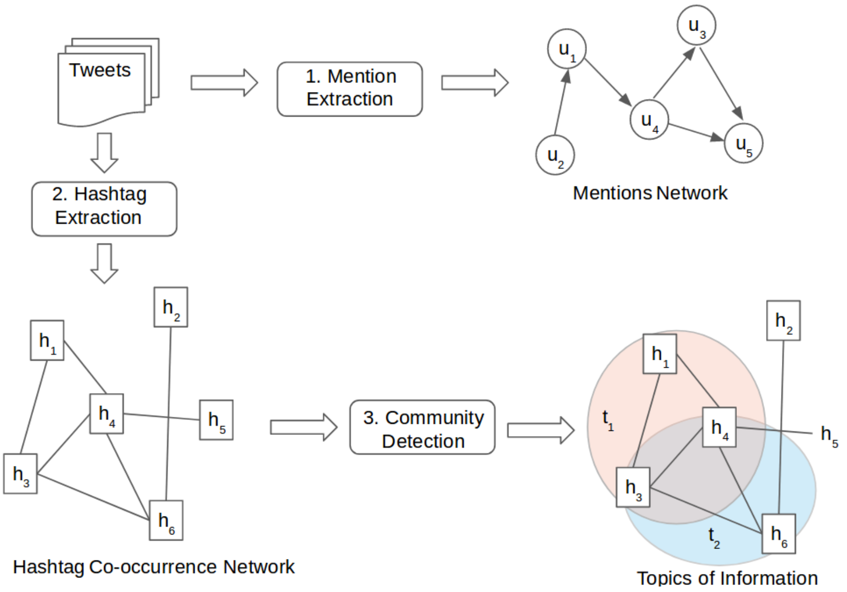


Figure 3.3: Implementation scheme for the basic networks preparation in three steps.

In the tweets collection phase, the tweets were collected and stored in the JSON document format[1], with annotations for the mentions and hashtags in each tweet. In step 1, we extracted the mentions and built a mentions network, by assigning a directed edge from the mentioning user to the mentioned user. We also applied two filters, one to select users with significant activity, which are those with 9 or more hashtags, and another to exclude users with less than three followees.

In step 2, we built a hashtag co-occurrence network, counting their join occurrences over all the collected tweets. The resulting network is undirected and each pair of hashtags is linked with positive weight, equal to the counted pairwise occurrences. We chose to omit this weight in Figure 3.3 to mantain a clean visualization. We filtered out every edge with weight of one or two, to discard hashtags that could have co-occurred randomly and reduce the noise in co-occurrences.

In step 3, we ran the community detection implementation of the Order Statistics Local Optimization Method[2] found in [39], setting it to iterate along with the infomap method [54] two times and Louvain's method [10] one time. The output was formed by the best modules from the three methods, without assigning nodes isolated from modules - singletons - to any communities.

It is important to highlight that we worked a single co-occurrence network and a single mentions network, both **static** and being formed from the accumulated tweet activity throughout the entire observation time window of the dataset, comprising 45 days. We assumed that hashtag associations do not suffer much seasonal effects in short periods of time and therefore remains relevant though the time measured. The mentions network

[1] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.htm
[2] http://www.oslom.org/

works as a proxy for the interaction, so it is plausible to look at the global representation from accumulated interactions.

3.2.3 Adoption Networks Preparation

To prepare the dataset for the computations prescribed by our model in Subsection 3.1.3, we discretized the time in our observation window. Notice that this phase is no longer static, as each different discretization leads different possible adoption events and a different baseline. We further summarize this phase.

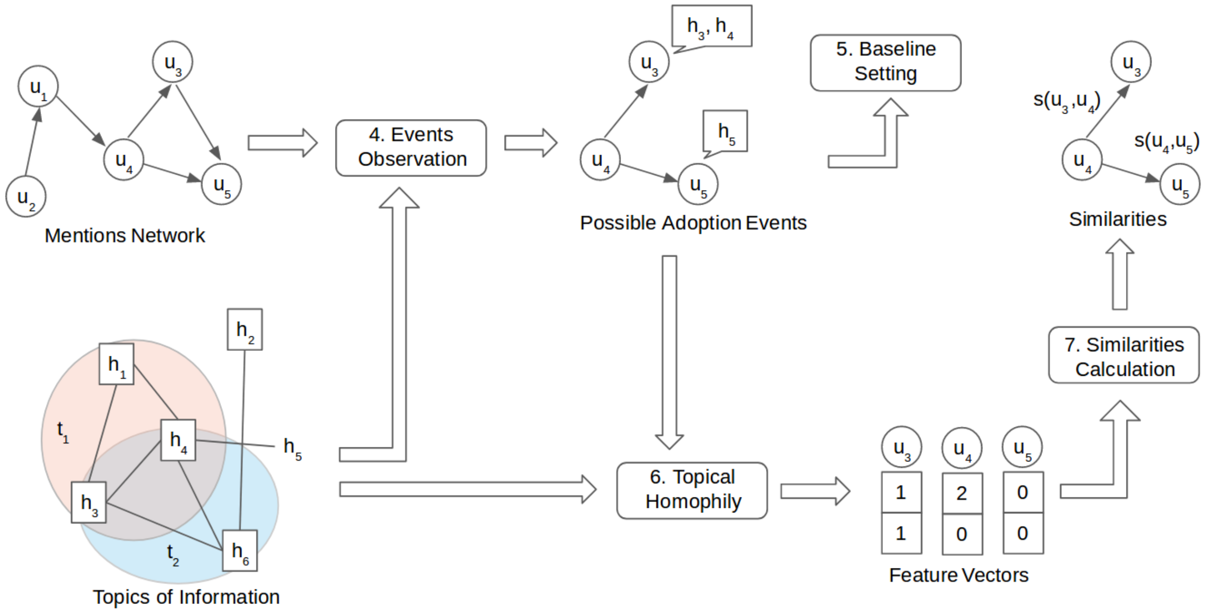


Figure 3.4: Implementation scheme for the simulations preparation in three steps.

Figure 3.4 departs from the two graphs produced in the Basic Networks Preparation phase, presented in Figure 3.3.

At step 4, we chose to set a time step size of one day. It means that we have a discrete set $T = \{1, 2, ..., 45\}$ and then, all the tweets are assigned to the relative day that they occurred. We produced this discretization by mapping each tweet of timestamp between 00:00:00 and 23:59:59 to the corresponding day, starting by $t = 1$ at September 18th of 2017. For example, a tweet with timestamp 22:56:47 of September 23th is mapped to $t = 6$. Now, at every time t, we observe a number of possible adoption events. We group each of them by the user agent, i.e., the user which received hashtags elligible for adoption. In Figure 3.4, we show the example of user u_4. We produced a graph of possible adoption events with each user pointing to all his neighbours and the hashtags used by his neighbours at time $t - 1$.

In order to validate our simulations, we set a baseline of adoptions in step 5. We considered an actual adoption whenever user u used a hashtag h in one of his messages at time t and this hashtag h figures in a possible adoption event of the same time t. In other words, if his neighbours used h at time $t-1$ and he uses h at time t, we count it as an actual adoption in our baseline. The baseline works as a reference for real meme adoptions extracted from our dataset.

Moving to step 6, we characterize the homophily among users through the topics that they participate. We built feature vectors to represent the topic interests of each user, as defined in Expression 3.1. The feature vector of each user agent is built from every hashtag from time 1 to $t-1$.

From these feature vectors, we can determine in step 7 how similar the users are, by taking the similarity defined in Expression 3.2. At the end of this process, we have all the possible adoption events, all the candidates for adoption at each situation, and all the similarities between the users that we are interested in. In the next section, we outline how we use these elements to simulate the adoption of memes and how the results of these simulations are used to validate our model.

3.3 Computational Simulation

Our objective is to simulate a meme adoption process that leads to the baseline of adoptions. At time t and given a number α of memory slots, let us consider that u emptied his memory slots and all the memes of the possible adoption events may fit his memory slots with some chance. If the user in time t posted a hashtag selected from the possible adoption events, the simulation considers he adopted it.

To evaluate the influence of the topical homophily in the meme adoption and whether it improves the adoption prediction, we performed two different kinds of simulation. The first one with chances equal to the λ function built from the topical homophily, i.e., the values given by the normalization of Expression 3.5. The second one have uniformly random chances. This is depicted in Figure 3.5.

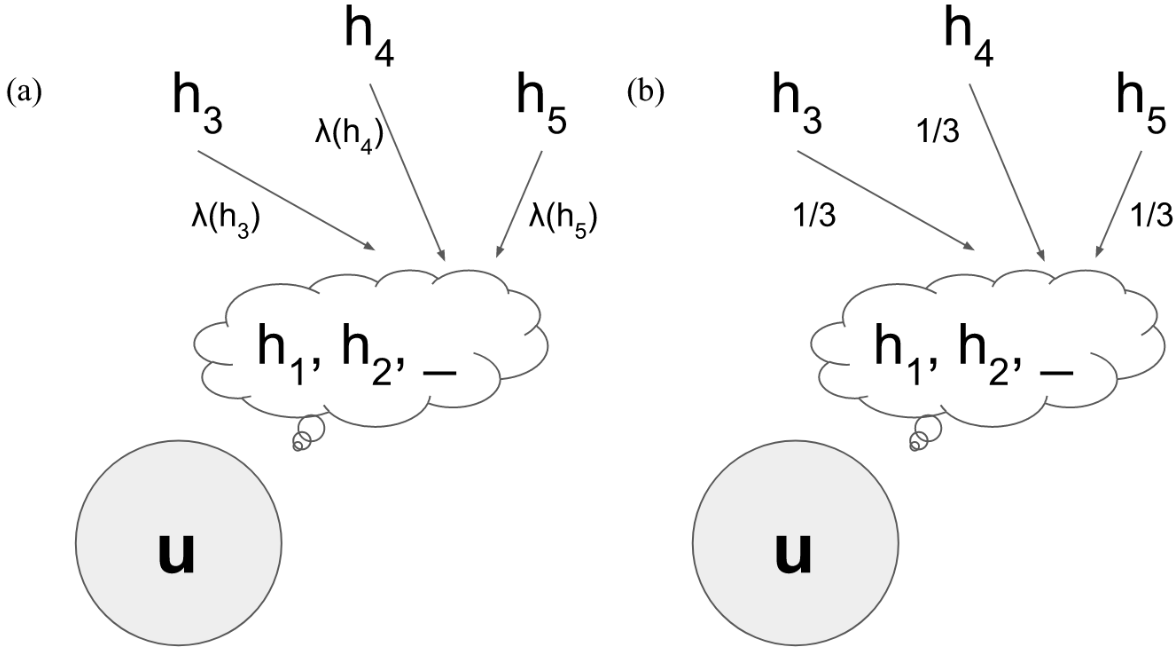


Figure 3.5: Simulation of adoptions. Either memes can be remembered with chances given by (a) the λ function from topical homophily or (b) equal chances.

We want to observe whether the memory configuration oriented by topical homophily is closer to the baseline than a random scheme of memory update. If the memes drawn in uniformly random chances provide a better or equal approximation to the baseline than our proposal, the topical homophily is not a good explanation to what is behind the selection mechanism in their competition for attention.

Therefore, we established different criteria to measure how distant these two simulation schemes are to the baseline.

3.4 Validation

We checked three properties regarding each set of adoptions generated by simulation and the adoptions from the baseline. They are the Kendall rank correlation coefficient, the meme lifespan precision and the third one we dubbed simply as precision.

By studying different values of α, we expected to obtain higher values of each metric to the simulations based on the topical homophily than those with uniformly random drawing. This would be an indication that topical homophily yields a closer process than the random selection.

3.5 Kendall rank correlation coefficient

To the first criterion, we summarized the adoptions of the baseline and the simulation by ranking the memes with their number of adoptions. We compare both with the Kendall rank correlation coefficient. This number, also known as Kendall tau metric [48], measures how much two ordinal associations in a set of variables are close. The Kendall τ coefficient is defined as:

$$\tau = \frac{(\#\, concordant\, pairs) \;-\; (\#\, discordant\, pairs)}{n(n-1)/2} \tag{3.7}$$

If we know that there is a strict total order relation in the ranked elements, it is possible to determine if the order of each pair of elements concord or discord between the two ranks. This number then goes from -1, when there is no pair of elements with the same order in the two ranks, to 1, when the ranks are exactly the same.

We chose this metric to validate our results as it is responsive to fluctuations in the number of adoptions. While the lowest and highest values will probably appear in similar positions in the two adoption ranks, there is a large space for discrepancies in the intermediate numbers of adoptions.

3.5.1 Meme Lifespan Precision

The second validation metric is based on the meme lifespan. The lifespan of a meme is defined as the longest time period in which the meme kept being adopted. In our choice of discretization, this value peaks at 45, if the meme was shared regularly through the entire observation window, and has a minimum of 1 if this meme could not last for more than one day in the social network.

The meme lifespan precision is built from both the simulated adoptions and the baseline adoptions. It is the fraction of memes with the same lifespan in both over the total number of memes. That is, if the simulation has a null meme lifespan precision, no meme lived as long as the memes in baseline lived. Notice that this case is impossible, due to the way we consider the simulated adoptions.

On the other hand, a simulation with meme lifespan precision equal to 1 prescribes memes which survive the exact same amount of the time as the baseline. Even though this number is not directly bonded with the popularity of a meme – memes may spread in a burst, reach the entire network and die fast –, it is an interesting metric because it is sensitive to time. A single miss of simulated adoption will rule the meme out of the correct prescriptions, and the chance of missing increases with the lifespan in the baseline.

3.5.2 Precision

This measure is obtained from simply counting how many adoptions happened in both the simulation and the baseline and dividing this amount by the total of baseline adoptions. We looked at this fraction because it is a clear calculation related to the two previous validation criteria.

Chapter 4

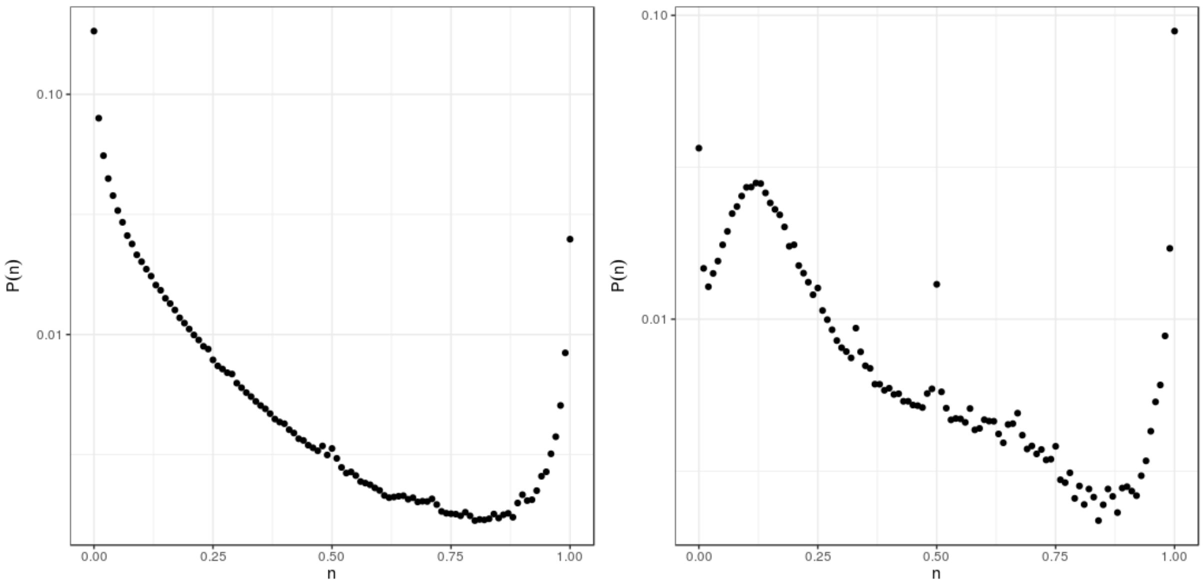
Results

In this chapter, we bring together the most important results in this work. Starting by Section 4.1, we reproduced the related work to check whether the similarities in our social network have a similar behavior to the literature. Section 4.2 shows the product of our discretization choice, in terms of possible adoption events. Section 4.3 is dedicated to describe the baseline of adoptions we obtained and what the simulations are aiming for. Finally, Section 4.4 compiles results of the simulations of our model.

A deeper description of the dataset is encountered in Appendix A, provided for supplementary material. We display frequency distribution plots by using a notation of 'n' to the observed value in the x-axis and 'P(n)' to the fraction of observations corresponding to 'n' in a semi-log scale y-axis. Unless expressly stated, all plots generated condense values in 100 bins to improve visualization. We decided to show most of the results this way because the majority of the distributions have high skewness, so a default histogram would have most of the values concentrated in the first bars, near the zero of the x-axis.

4.1 Topical Homophily

Based on a similar approach from Cardoso et al. [13] in our first analysis, we wanted to see the distribution of similarities over the set of all 261,521 users, selected by being considered active - i.e., those with 9 or more hashtags - as well as the same values averaged over the neighbours of each user. The plots are in Figure 4.1.



(a) Similarity measures.

(b) Average similarities over each user.

Figure 4.1: Distribution of the calculated similarity values.

In Figure 4.1a, n represents a similarity measure between each pair of neighbours in the entire mentions network. P(n) represents the fraction of users that bears the respective similarity. As the chart shows, most of the connections have small similarities and the fraction decreases for higher similarities.

Figure 4.1b has the same kind of visualization for the average values of similarities over each user's neighbours. The similarity is calculated so that it is 0 if we are comparing two social network users who share no topic of information in their messages, as opposed to 1 if the pair of users talks about the exact same topics. There is a continuous and steep increase starting in similarities values nearly 0.9. We started a small investigation to understand what could cause a significant number of users with identical topic interests with all their neighbours and, but we could not reach noteworthy results. This is left to Appendix B.

4.2 Possible Adoption Events

In this section, we specify the set of possible adoption events found in our dataset for the time step size of 1 day, as we mentioned earlier in Subsection 3.1.4 of Chapter 3. The 17,239,972 possible adoption events retrieved formed a highly heterogeneous set with 115,054 memes and 106,994 users. Therefore, we opted to plot how they are distributed over the 45 time points, instead of the semi-log format specified at the beginning of this chapter. The chart in Figure 4.2 shows the daily quantity of possible adoption events in blue in opposition to actual adoptions in yellow. In the average of all observed days, nearly 6% of the possible adoption events resulted in actual adoptions. Each possible adoption event can be mapped to a hashtag eligible for adoption through the criteria defined in our methodology.

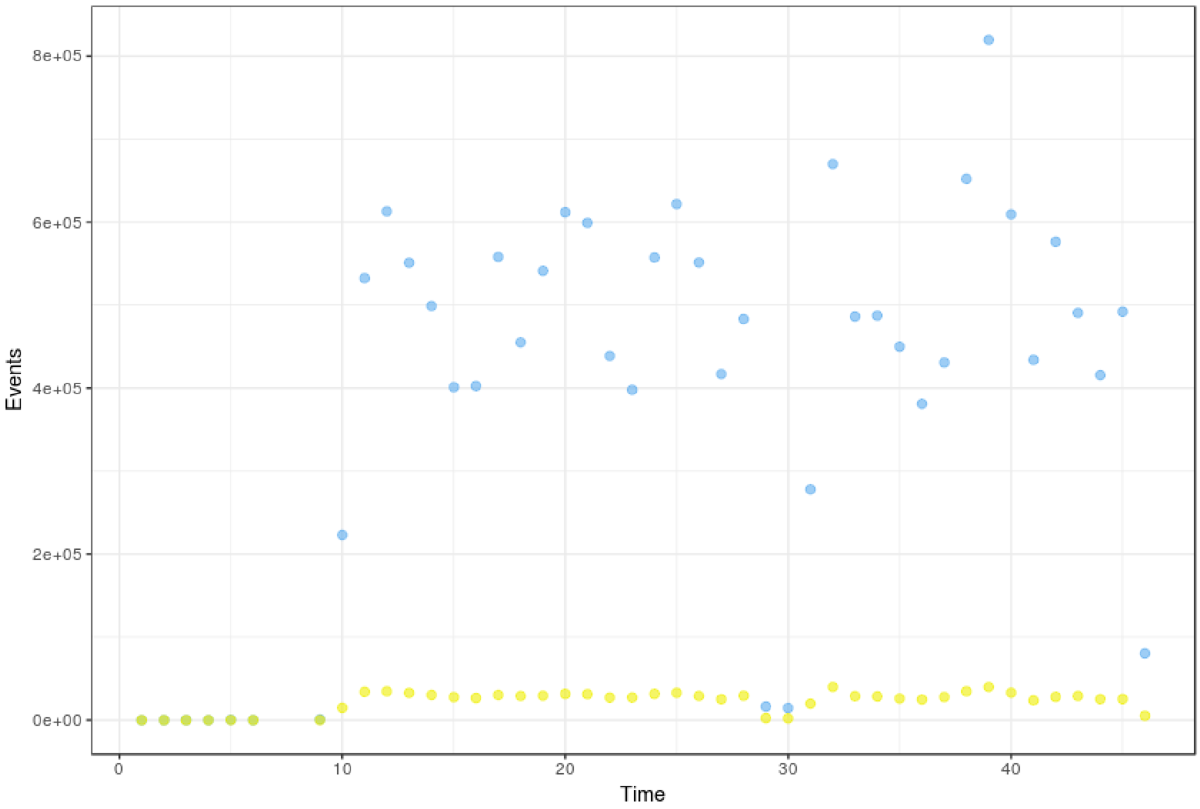
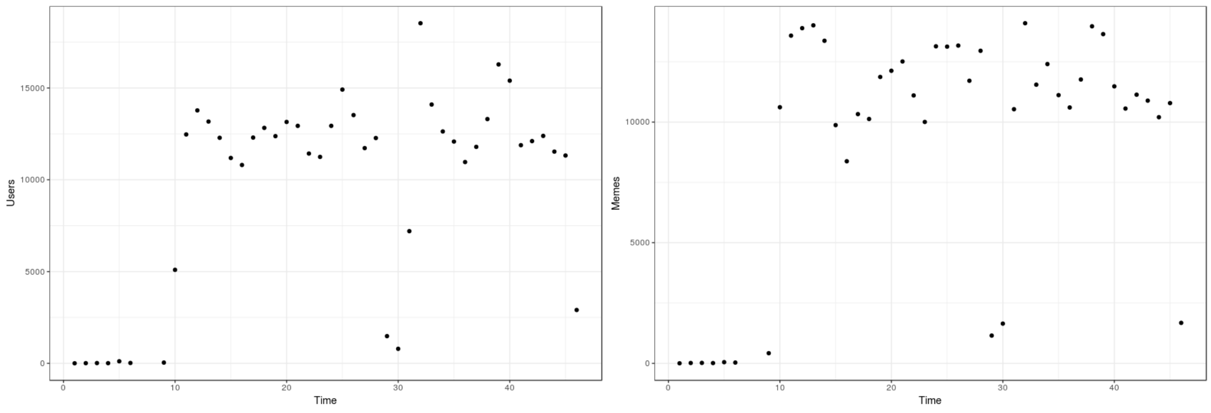


Figure 4.2: Observed quantity of possible adoption events in blue and actual adoptions in yellow for each time value.

In the ten first days, the values are regularly low. We think this may be caused due to problems in the first period of the collection of the data. We can see that, after day 10 onwards, the number of events shows high fluctuation, what is expected from such a dynamical system as the social media. The actual adoptions have much less variation in comparison to the possible ones. The next two charts carry similar information but counting how many agent users and how many memes were implicated with possible adoption events in the same period.



(a) Different users.

(b) Different memes.

Figure 4.3: Quantity of users (a) and memes (b) related in the possible adoption events over time.

At each possible adoption event, we found its topical fitness values defined by the Expression 3.5. They are distributed as Figure 4.4 displays.

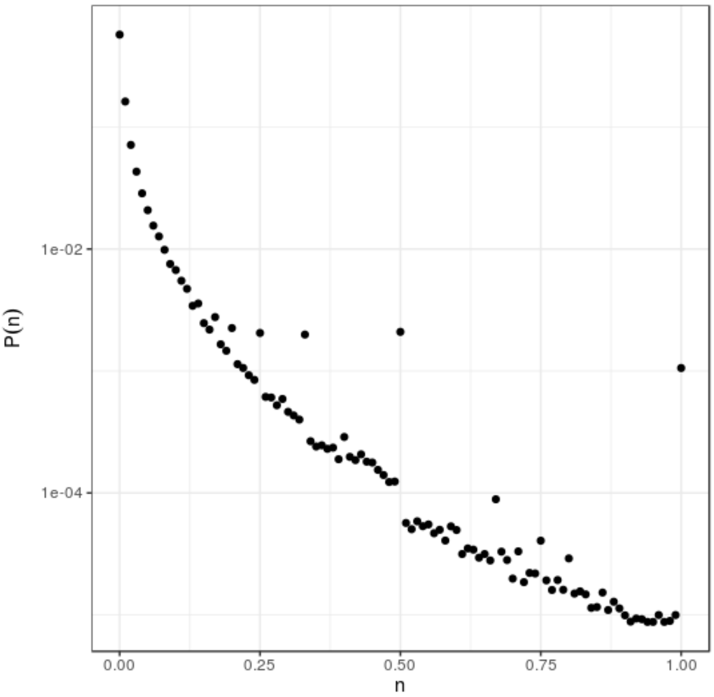
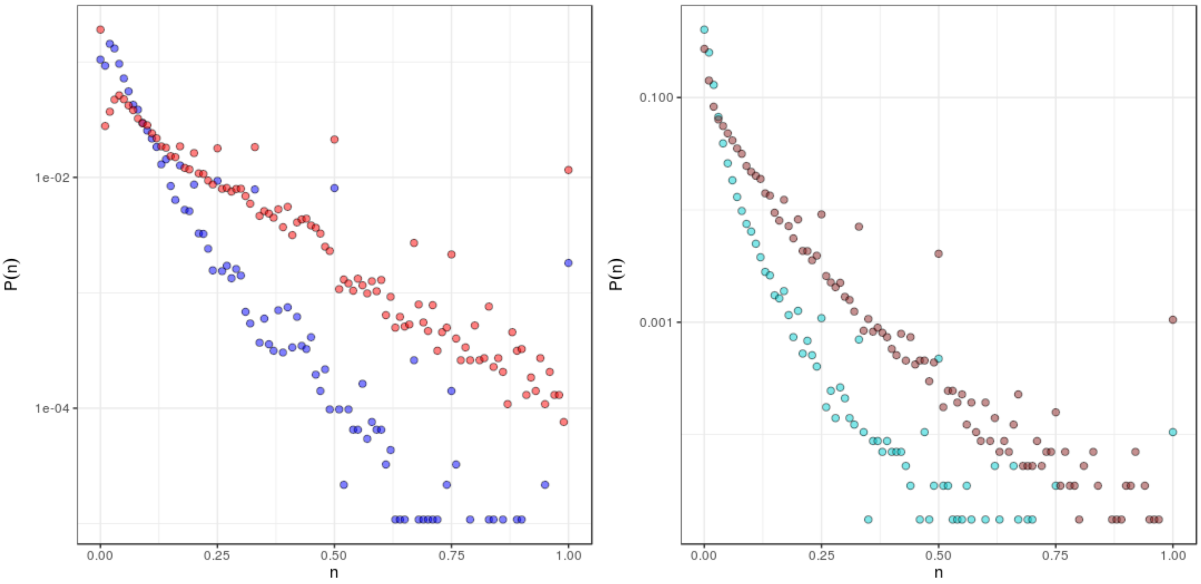


Figure 4.4: Distribution of the calculated topical fitnesses.

We remind that the topical fitness of each meme is a fraction of the similarities to every neighbour, so n goes from 0, when the neighbours posting the meme have null similarity, to 1 when every neighbour with positive similarity posted the meme. P(n) represents the fraction of topical fitnesses taking the value of n.

The values exhibit an interesting behavior of a high specialization of topical interests - i.e., most of the adoption events are similar to few topics -, aside from outliers between $P(n) =$ 1e-2 and $P(n) =$ 1e-3. This wide variation of hashtag attractiveness may suggest a greater difference later in our simulations when differentiating topical homophily adoption from the uniformly random one. The outliers jumps from fractions which appear to be 1/4, 1/3, 1/2 and 1, which would reflect people with very few options of adoption and possibly too similar neighbours. This explanation is consistent with the degree distribution of the mentions network and the similarity values in Figure 4.1b.

We averaged the topical fitnesses values for each meme and for each user to improve the visualization between the distribution of adopted and non-adopted memes.



(a) Averaged over users

(b) Averaged over hashtags

Figure 4.5: Distribution of average topical fitnesses of adopted memes (red/brown) and non-adopted memes (blue/cyan).

It is possible to see that the values of fitness for adopted memes are slightly more distributed to the right, meaning they average at higher topical fitnesses than the correspondent for non-adopted. This is a good indicator if we want to use this measure to predict adoption.

When the λ function is derived from a topical fitness equal to 1 does not necessarily mean that a meme is adopted for sure. For instance, in the event that all memes eligible for adoption have fitness equal to one, the selection using λ is the same as the uniformly random. After analyzing the adoptions in the timeline and computing the values for a λ function, we further inspect the baseline.

4.3 Baseline Characterization

In this section, we analyze the baseline collected for our observation of actual adoptions, a procedure elucidated in step 5 of Subsection 3.2.3. We gathered information on the behavior of the baseline extracted from our dataset to clarify the challenge of simulating it. We initiate with Table 4.1 by showing the most adopted memes, with their designating hashtags, the number of possible adoption events they figure in, the total number of adoptions they amassed over the observation window, and the ratio between adoptions and events.

Table 4.1: The 10 most adopted memes in the baseline.

Hashtag	Events	Adoptions	Ratio
MPN	24,422	17,515	0.7172
Venezuela	78,764	6,754	0.0857
BTSShow	9,434	6,257	0.6632
Catalunya	57,450	5,105	0.0889
LulaPorMinasGerais	26,023	4,891	0.1879
GloboLixo	26,638	4,735	0.1778
Bolsonaro2018	40,196	4,696	0.1168
Política	56,393	4,141	0.0734
1Oct	43,441	3,834	0.0883
1O	40,704	3,707	0.0911

Even though these memes are successful by virtue of the number of adoptions - our closest metric to the popularity or cascade size [30] - the ratio between how many times they are adopted and how many times they are exposed is generally low, with four of the memes having a ratio of less than 10%. We looked at the sum of adoptions over the observation period to all memes in the baseline and examined the distribution of these values in Figure 4.6.

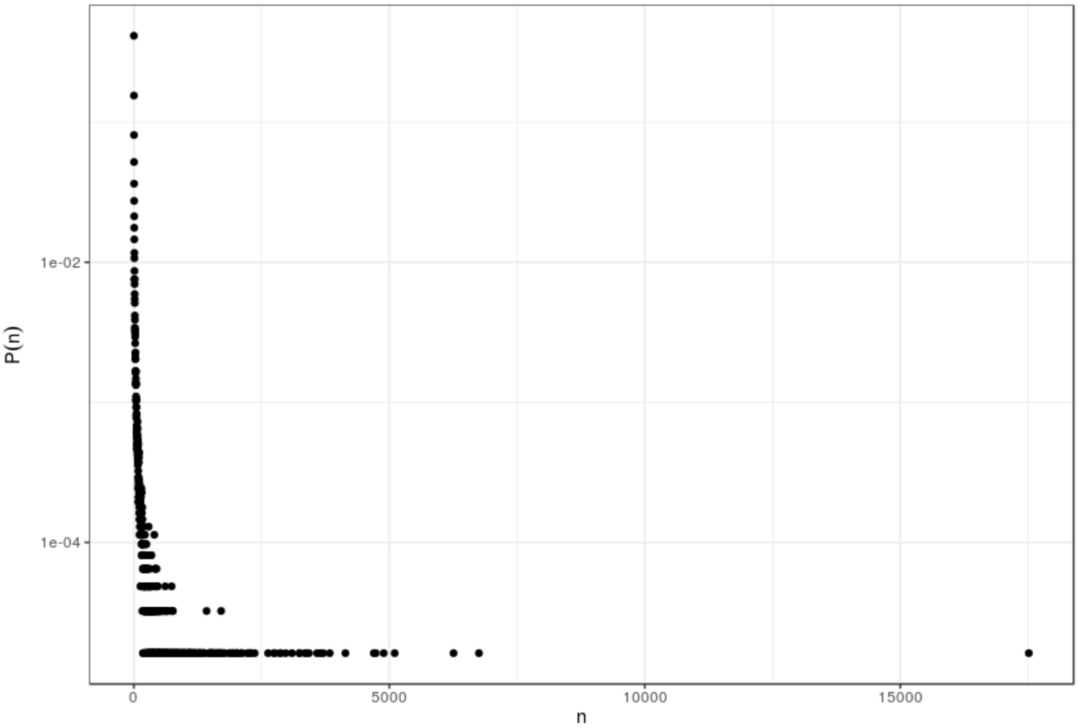


Figure 4.6: Distribution of the number of adoptions of each meme in the baseline.

This is a very highly skewed distribution, an evidence to the adversity of simulating it. It is possible to see that the 10 first memes in the rank of adoptions populate nearly 4/5 of the axis to measured values of the number of adoptions. Even the semi-log representation hardly eases the concentration of values near 0.

We further checked another measure, the lifespan of memes. This quantity is observed from the longest period of time in days that the meme kept being adopted. We built a table analogous to Table 4.1, but this time with data grouped by meme lifespan. The second column has the number of different memes which achieved the lifespan, the third accounts for the sum of their adoptions, and the fourth the average number of adoptions to memes in this level.

Table 4.2: The top 10 longest meme lifespans in the baseline.

Lifespan	# Memes	Total Adoptions	Avg # Adoptions
38	10	15,177	1,517.70
37	45	96,625	2,147.22
36	13	9,422	724.77
35	2	1,106	553.00
33	1	1,23	123.00
31	3	1,701	567.00
29	3	2,755	918.33
28	2	1,965	982.50
27	3	2,363	787.66
25	2	1,467	733.50

The first thing to draw attention is that the memes indeed face a fierce competition, as only 84 of the 115,054 achieved the top lifespans. More than half lived to near the maximum observed for our baseline, but they had a very good performance of adoptions in average. We present the distribution of lifespans in Figure 4.7.

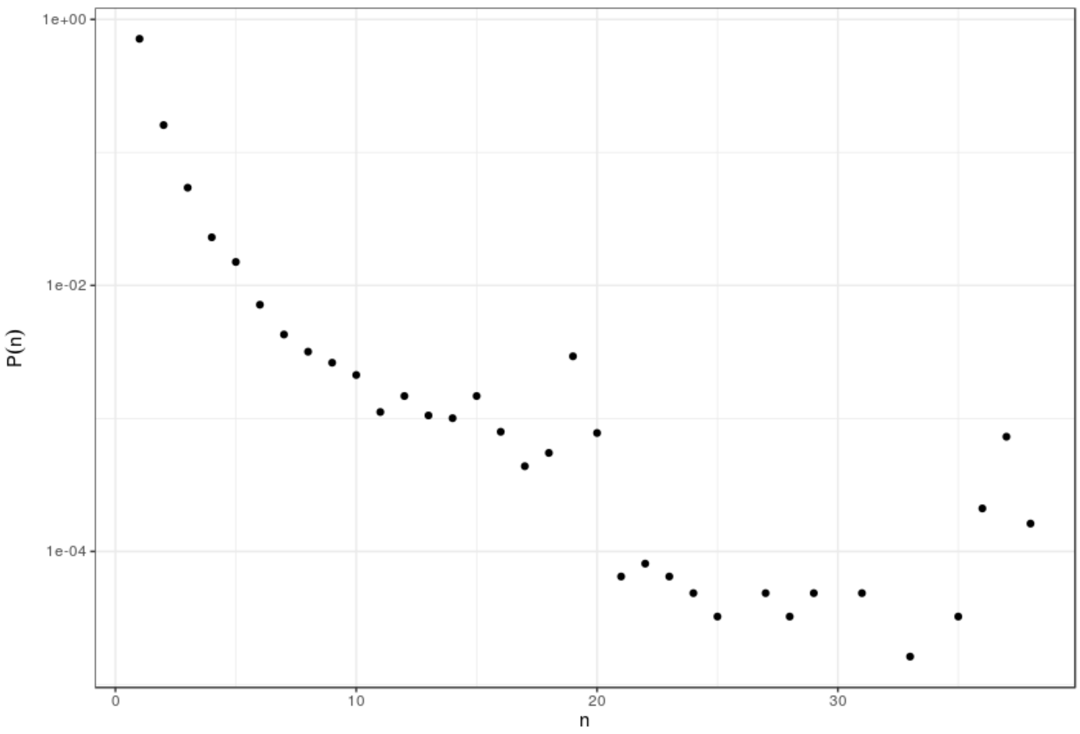


Figure 4.7: Adoptions counted to each meme over its lifespan.

Albeit the lifespans have a small range of values, the semi-log axis still help us to visualize where the memes concentrate their duration. We can spot that past 20 days, it is difficult to survive. We also offer the relation between the number of adoptions and the lifespan of memes in our dataset in Figure 4.8.

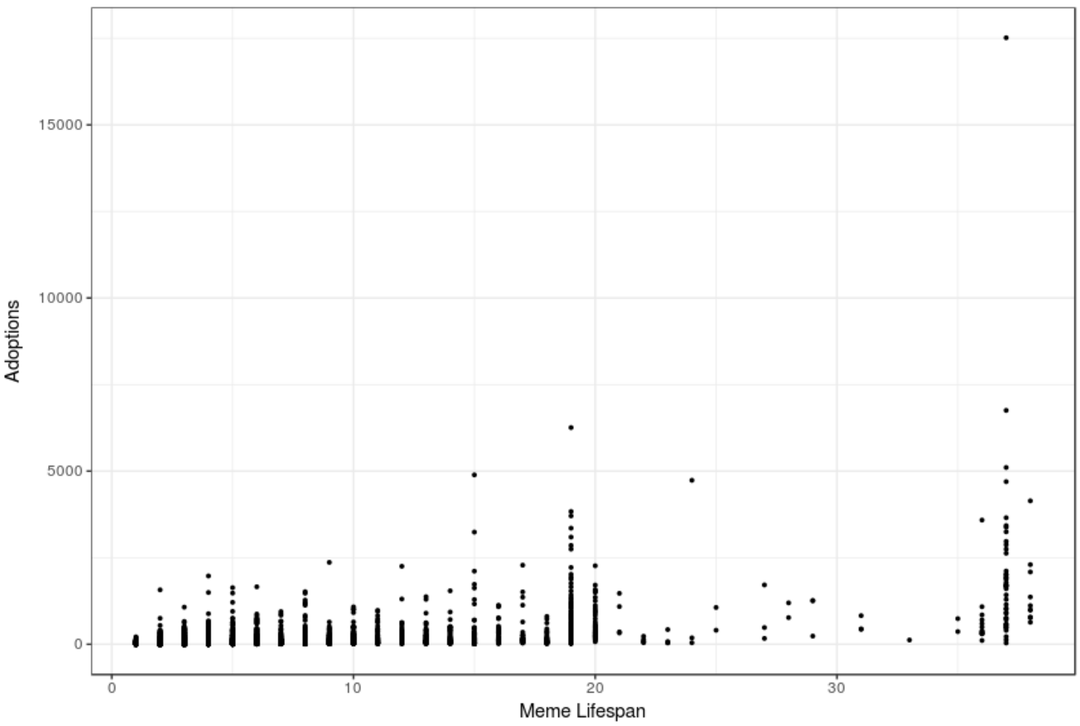


Figure 4.8: Adoptions counted to each meme over its lifespan.

We notice that memes with a given lifespan value are poorly represented by their average number of adoptions because the distribution is largely spread across the y-axis. This fact reinforces the heterogeneous character of the baseline, as the relationship between the lifespan of a meme and its number of adoptions is not direct. Now that we know better our baseline, we proceed with our plan to simulate it.

4.4 Simulations

In this section, we visualize the three kinds of validation described in Section 3.4 which are yielded by the comparison of our simulations with the baseline. We performed five simulations of two types, topical homophily simulations and uniformly random simulations. Each of the ten simulations was executed for five different values α of memory slots, which are $\alpha = 5$, $\alpha = 10$, $\alpha = 20$, $\alpha = 30$ and $\alpha = 40$, in a total of fifty simulations. The results are demonstrated in a standardized structure of two tables with values yielded by the two simulation types and a summarizing table with average numbers, standard deviations and differences between the averages. Finally, a line chart presents a visual representation of the third table.

We are interested in recognizing a consistent difference between the topical homophily and the uniformly random scheme. The difference is appreciable when it is significantly higher than the standard deviations of the validation metric of the simulations, and it is consistent if it remains appreciable for a varying number α of memory slots.

4.4.1 Kendall Rank Correlation

As described in Subsection 3.5, we built an adoptions rank by counting how many adoptions each meme had and sorting them in decreasing order. Then, these relations are compared with the Kendall rank correlation coefficient implemented in the R package 'Kendall' [1]. The resulting comparisons between the baseline rank and the ranks obtained from simulated adoptions are presented at Table 4.3, for the topical homophily adoption scheme, and Table 4.4, for uniformly random adoptions. Both situations are summarized in Table 4.5 and Figure 4.9. All the p-values for the computed correlations are near the tolerance of the floating point in R, which is 2.2e-16.

Table 4.3: Kendall τ coefficient values obtained between the topical homophily simulated adoptions rank and the baseline adoptions rank.

Run	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 40$
1	0.09188175	0.14193400	0.18832900	0.18592270	0.23649690
2	0.09169501	0.14260240	0.18902960	0.18600810	0.23633970
3	0.09151094	0.14349660	0.18833820	0.18620130	0.23578900
4	0.09191245	0.14266950	0.18879450	0.18685550	0.23722210
5	0.09125654	0.14320110	0.18820800	0.18583570	0.23612200

[1] https://cran.r-project.org/web/packages/Kendall/Kendall.pdf

Table 4.4: Kendall τ coefficient values obtained between the uniformly random simulated adoptions rank and the baseline adoptions rank.

Run	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 40$
1	0.08140281	0.12912550	0.17468640	0.16814190	0.22306760
2	0.08127601	0.12929860	0.17413340	0.16867890	0.22407270
3	0.08203281	0.12910290	0.17496260	0.16796950	0.22339410
4	0.08136679	0.12831900	0.17589960	0.16888870	0.22462950
5	0.08066648	0.12864280	0.17586120	0.16774080	0.22339700

Table 4.5: Summarized Kendall τ coefficient results between both types of simulations. 'Avg' stands for average, 'SD' for standard deviation, 'TH' discriminates the topical homophily simulations and 'UR' the uniformly random simulations, and the difference is taken between the average values.

	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 40$
Avg TH τ	0.09165134	0.14278072	0.18853986	0.18616466	0.23639394
SD TH τ	0.00027329	0.00060214	0.00035353	0.00040923	0.00053369
Avg UR τ	0.08134898	0.12889776	0.17510864	0.16828396	0.22371218
SD UR τ	0.00048528	0.00040453	0.00076529	0.00048368	0.00062997
Difference	0.01030236	0.01388296	0.01343122	0.01788070	0.01268176

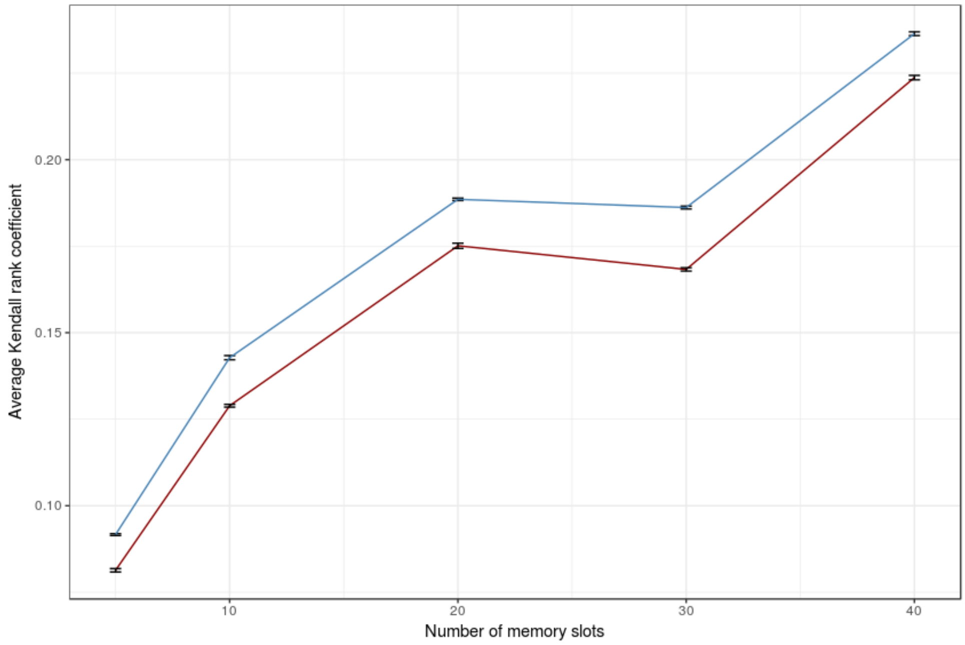


Figure 4.9: Average Kendall τ and their standard deviations obtained for topical homophily simulations in the upper blue line and uniformly random simulations in the lower red line, over the number α of memory slots.

4.4.2 Meme Lifespan Precision

The next comparison is based on the observed meme lifespan, a validation metric defined in Subsection 3.5.1. Again, the meme lifespan is given by the biggest period of consecutive days the meme appeared in adoptions. For each lifespan value, we count the portion of memes which had their simulated adoptions duration equal to those in the baseline.

In the same fashion as the previous subsection, Table 4.6 has the meme lifespan precisions for each simulation in the topical homophily scheme, Table 4.7 brings meme lifespan precisions of the uniformly random simulations and Table 4.8 summarizes them along with Figure 4.10.

Table 4.6: Meme lifespan precision values obtained between the topical homophily simulated adoptions and the baseline adoptions.

Run	Alpha=5	Alpha=10	Alpha=20	Alpha=30	Alpha=40
1	0.54531614	0.65371534	0.75207307	0.75621922	0.84004923
2	0.54489505	0.65510818	0.75239699	0.75756348	0.84011401
3	0.54521896	0.65577222	0.75251036	0.75701282	0.83904508
4	0.54576962	0.65556167	0.75267232	0.75691565	0.84106957
5	0.54670899	0.65552928	0.75288287	0.75549041	0.84009782

Table 4.7: Meme lifespan precision values obtained between the uniformly random simulated adoptions and the baseline adoptions.

Run	Alpha=5	Alpha=10	Alpha=20	Alpha=30	Alpha=40
1	0.51721624	0.61963915	0.72319577	0.71987561	0.81457307
2	0.51961324	0.62205234	0.72147901	0.72327675	0.81518852
3	0.52053640	0.61989828	0.72089595	0.72121987	0.81429774
4	0.51783169	0.61941241	0.72121987	0.72259652	0.81541526
5	0.51694091	0.62080526	0.72353589	0.72131705	0.81415198

Table 4.8: Summarized meme lifespan precision results between both types of simulations. 'Avg' stands for average, 'SD' for standard deviation, 'TH' discriminates the topical homophily simulations and 'UR' the uniformly random simulations, and the difference is taken between the average values.

	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 40$
AVG TH	0.54558175	0.65513734	0.75250712	0.75664032	0.84007514
SD TH	0.00070350	0.00083061	0.00030373	0.00080100	0.00071634
AVG UR	0.51842770	0.62036149	0.72206530	0.72165716	0.81472531
SD UR	0.00157206	0.00108316	0.00121103	0.00132157	0.00055346
Difference	0.02715405	0.03477584	0.03044182	0.03498315	0.02534983

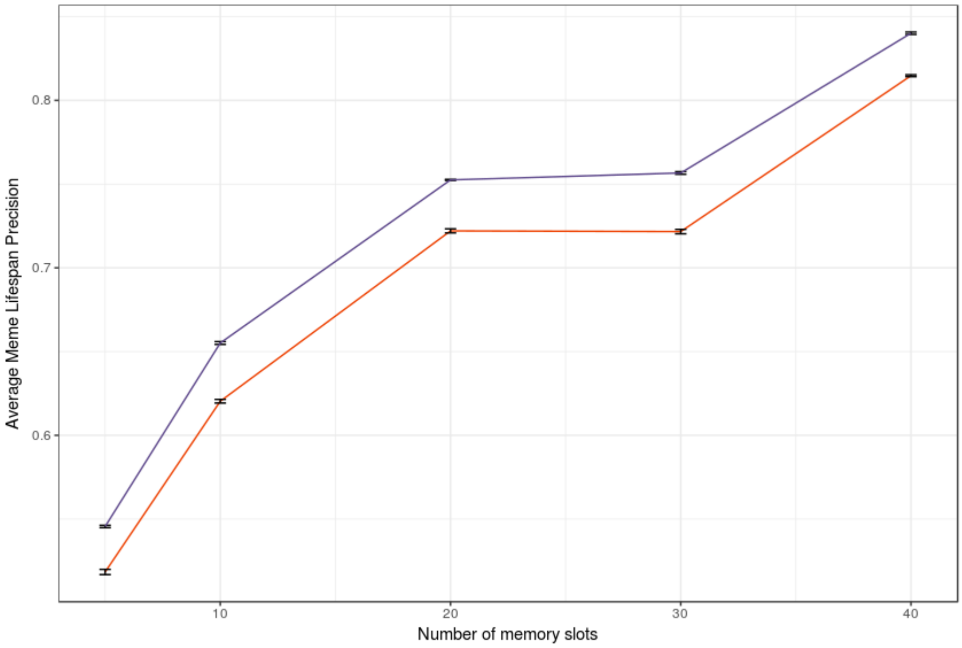


Figure 4.10: Average meme lifespan precisions and their standard deviations obtained for topical homophily simulations in the upper purple line and uniformly random simulations in the lower orange line, over the number α of memory slots.

4.4.3 Precision

Our last comparison evaluates the rate of correctly prescribed adoptions of the simulations, as stated in Subsection 3.5.2. Figure 4.11 is derived from Table 4.11, which is assembled from the topical homophily scheme results in Table 4.9 and the uniformly random scheme in Table 4.10.

Table 4.9: Precision values obtained between the topical homophily simulated adoptions and the baseline adoptions.

Run	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 40$
1	0.52542141	0.65142868	0.784435299	0.81445040	0.89633557
2	0.52552130	0.65080939	0.784621402	0.81436628	0.89634608
3	0.52559069	0.65173465	0.784506796	0.81436208	0.89673721
4	0.52585670	0.65137821	0.784468945	0.81473323	0.89666151
5	0.52551394	0.65139083	0.784703413	0.81418754	0.89655742

Table 4.10: Precision values obtained between the uniformly random simulated adoptions and the baseline adoptions.

Run	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 40$
1	0.45520499	0.56710525	0.70188878	0.72351349	0.83426139
2	0.45552147	0.56691810	0.70171950	0.72300776	0.83441070
3	0.45560873	0.56704427	0.70184042	0.72325484	0.83435602
4	0.45518396	0.56696016	0.70171530	0.72362705	0.83448955
5	0.45520393	0.56715993	0.70193505	0.72352401	0.83430870

Table 4.11: Summarized precision results between both types of simulations. ’Avg’ stands for average, ’SD’ for standard deviation, ’TH’ discriminates the topical homophily simulations and ’UR’ the uniformly random simulations, and the difference is taken between the average values.

	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 40$
Avg TH	0.52558080	0.65134840	0.78454720	0.81441990	0.89652760
SD TH	0.00016556	0.00033498	0.00011201	0.00019957	0.00018206
Avg UR	0.45534460	0.56703750	0.70181980	0.72338540	0.83436530
SD UR	0.00020378	0.00009983	0.00009930	0.00025183	0.00008886
Difference	0.07023620	0.08431090	0.08272740	0.09103450	0.06216230

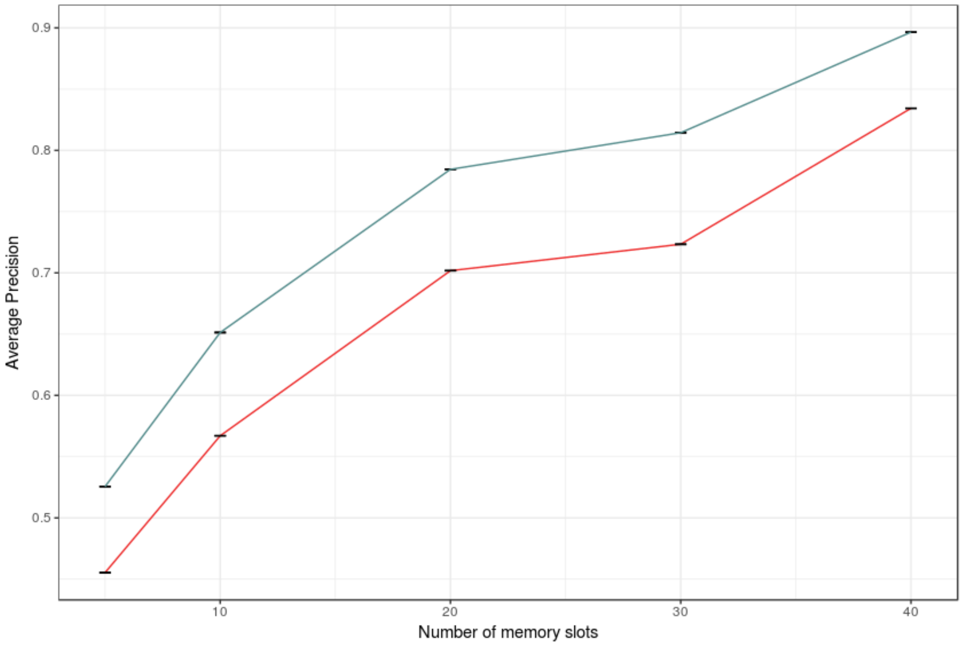


Figure 4.11: Average precisions and their standard deviations obtained for topical homophily simulations in the upper blue line and uniformly random simulations in the lower red line, over the number α of memory slots.

We can see the number α of memory slots as a proxy for the selective pressure faced by memes, in the sense that the lower this value is, the more thriving is required for memes to survive. More than a relation between our predictions and α, through the three presented results we wanted to observe a consistent difference between random selection and homophily selection across different settings of selective pressure.

Based on the results of our experiments, we argue that our most sensitive result in the difference of simulations generated by topical homophily and uniformly randomness is that obtained from the Kendall τ coefficient. Even though the meme lifespan precision is a pertinent dimension of reproduction, the range of values it can assume is rather small. The true positive precision has a limited extent to the behavior of the system in a manner of binary classification, without the coupling of a recall measure. This recall measure requires us to produce false negatives, which, in this context, are simulated through the production of original memes, which is out of the scope of this research. As for the Kendall rank correlation, we saw in Figure 4.6 that the adoptions are distributed in a very skewed manner. It would probably be easier to approximate the lowest and the highest values of adoptions, but the in-between space is rather large and makes us prone to error, specially considering the pairwise combinations of order relations over memes are counted in over 6.6e+10.

Chapter 5

Conclusions

In order to conclude our work, Section 5.1 reviews the results of our methodology as a means to achieve a critic view of the entire work and discuss some aspects relevant to the statistical validity of our results; Section 5.2 summarizes the contributions of what we found through the present effort of scientific investigation; and Section 5.3 explain the possibilities these contributions carve for future scientific endeavors.

5.1 Discussion

As stated before by the work of Morstatter et al [47], Twitter's Streaming API samples nearly 1% of all tweets. This limitation is generally overcame by extending the duration of collection in the related work we examined, so that the sample size is reasonable. We judged that the size of our sample was enough to this study, but it does not necessarily means that the same methods carried out in a bigger dataset would not retrieve different findings. The same can be observed about the number of simulations for each α and to a grainer range of its values. Likewise, a thinner time discretization may capture a unknown number of adoptions we could have missed.

The lack of statistical assessment on the impact of working with static topics of information pose an unknown danger to the validity of our experiment. As stated by the end of Subsection 3.2.2, we part from the supposition that hashtag association is not seasonal is such small time period, but it does not imply necessarily that the hashtags converge to some sort of steady state of association and, the greater the period we are looking, the stabler this consideration is. In fact, recent works are already presenting the debate over temporal dynamics of topics of informations [44]. We follow the trail of non-Markovian cascade models, so the extra caution with temporal evolution is supposed to be a scientific guideline. Yet both the simplification of a static topical configuration and daily update memory slots for meme adoption go in the direction of the Markovian property. We thought this would require a bigger research effort, so in order to narrow down the scope of this work to a master thesis, we decided nevertheless to adhere to this consideration.

Improvements could have been made by the planned application of statistical frameworks to bias detection in the data set, caused by our choice of keywords in the Twitter Streaming API. Most of the dataset filterings we proposed were based on related works, but they are not always objectively fundamented in statistical frameworks.

5.2 Contributions

We recall that this work leans over the role of homophily in the process of memes competing over attention in online social networks, as stated by our research question. For the sake of providing an answer to this, we worked with a specific characterization of homophily, based on a quantitative framework from topics of information. The state of art found in the literature does not differentiate how attractive each meme can be regarding each user in a local sense. We chose to work in this limitation.

Our main contribution in this work is a model for meme adoption which takes into account homophily measured by topics of information. To the best of our knowledge, this is a comparatively simple model which can be reproduced with clear steps of implementation, provided a dataset with few procedures of preparation.

Other contributions are the empirical evidences of meme competition and the effects of homophily in this context. We have seen that very few memes are able to survive through long periods and our characterization of homophily yields better prediction than uniformly random adoption of memes. Even though these discoveries are statistically questionable to a degree, our model supports a satisfiable reproducibility to carry on such questioning exercises.

Even though this is a initial study about memory behavior and memes in online social networks considering the scale of data, simulations and model complexity, we expect these contributions may corroborate generalizations of the other models, taking into account the heterogeneous attractiveness of meme adoptions to online social media users.

5.3 Future Works

The first extension to our work is to investigate memory models for meme adoption. This kind of memory can present a deeply diversified configuration, as the previous models that we studied have shown in the user activity. For instance, a social network user, with millions of followers and very few followees, acts much more like a broadcaster, always composing messages of his own, as opposed to profiles with lots of retweets and much less followers than followees. Both are different from a user with a balance between followers and followees in moderate numbers. The discerning of how our memory works in these systems may be the key to identify artificial social media profiles, one of the greatest open challenges in this field [27], which unleashed a contemporary technology race.

Another good opportunity is to separate what adoptions were caused by social influence rather the homophily, much like in the fashion of the research done by Aral et al. [2]. The interplay between these two properties should synthesize a robust model of meme fitness. We understand that our model can be enhanced to these research challenges in a near future.

The advancements on evolutionary modelling of memes may also be developed later into a real-time forecasting and population monitoring system of the meme competition environment, and lead to sophisticated social media with self-adjusted mechanisms for information curation.

Bibliography

[1] Lada A Adamic, Thomas M Lento, Eytan Adar, and Pauline C Ng. Information evolution in social networks. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 473–482. ACM, 2016.

[2] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[3] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, page 1215842, 2012.

[4] Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, 41(2):203–226, 1997.

[5] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[6] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[7] Raquel A Baños, Javier Borge-Holthoefer, and Yamir Moreno. The role of hidden influentials in the diffusion of online information cascades. *EPJ Data Science*, 2(1):6, 2013.

[8] Albert-László Barabási. *Network science*. Cambridge University Press, 2016.

[9] Samuel Barbosa, Roberto M Cesar-Jr, and Dan Cosley. Using text similarity to detect social interactions not captured by formal reply mechanisms. In *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pages 36–46. IEEE, 2015.

[10] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[11] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[12] Petko Bogdanov, Michael Busch, Jeff Moehlis, Ambuj K Singh, and Boleslaw K Szymanski. The social media genome: Modeling individual topic-specific behavior in social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 236–242. ACM, 2013.

[13] Felipe Maciel Cardoso, Sandro Meloni, Andre Santanche, and Yamir Moreno. Topical homophily in online social systems. *arXiv preprint arXiv:1707.06525*, 2017.

[14] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining*, page 4. ACM, 2010.

[15] Damon Centola, Juan Carlos Gonzalez-Avella, Victor M Eguiluz, and Maxi San Miguel. Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution*, 51(6):905–929, 2007.

[16] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[17] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, P Krishna Gummadi, et al. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30, 2010.

[18] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. ACM, 2014.

[19] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. The production of information in the attention economy. *Scientific reports*, 5:9452, 2015.

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

[21] GKD Crozier. Reconsidering cultural selection theory. *The British Journal for the Philosophy of Science*, 59(3):455–479, 2008.

[22] Richard Dawkins. *The selfish gene.* Oxford university press, 2016.

[23] Munmun De Choudhury. Tie formation on twitter: Homophily and structure of egocentric networks. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 465–470. IEEE, 2011.

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

[25] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world.* Cambridge University Press, 2010.

[26] Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. In *Advances in social networks analysis and mining (ASONAM), 2013 IEEE/ACM international conference on*, pages 548–555.

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

[28] James P Gleeson, Davide Cellai, Jukka-Pekka Onnela, Mason A Porter, and Felix Reed-Tsochas. A simple generative model of collective online behavior. *Proceedings of the National Academy of Sciences*, 111(29):10411–10415, 2014.

[29] James P Gleeson, Kevin P O'Sullivan, Raquel A Baños, and Yamir Moreno. Effects of network structure, competition and memory time on social spreading phenomena. *Physical Review X*, 6(2):021019, 2016.

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

[31] Michael H Goldhaber. The attention economy and the net. *First Monday*, 2(4), 1997.

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

[33] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[34] Tony Hey, Stewart Tansley, Kristin M Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[35] Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178. ACM, 2010.

[36] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. *Theory of Computing*, 11(4):105–147, 2015.

[37] Gueorgi Kossinets and Duncan J Watts. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450, 2009.

[38] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[39] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato. Finding statistically significant communities in networks. *PloS one*, 6(4):e18961, 2011.

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

[41] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[42] Changhyun Lee, Haewoon Kwak, Hosung Park, and Sue Moon. Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1137–1138, New York, NY, USA, 2010. ACM.

[43] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[44] Philipp Lorenz, Frederik Wolf, Jonas Braun, Nataša Djurdjevac Conrad, and Philipp Hövel. Capturing the dynamics of hashtag-communities. In Chantal Cherifi, Hocine Cherifi, Márton Karsai, and Mirco Musolesi, editors, *Complex Networks & Their Applications VI*, pages 401–413, Cham, 2018. Springer International Publishing.

[45] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

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

[47] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.

[48] RB Nelsen. Kendall tau metric. *Encyclopaedia of mathematics*, 3:226–227, 2001.

[49] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554, 2012.

[50] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[51] Xiaoyan Qiu, Diego FM Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1(7):0132, 2017.

[52] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM, 2002.

[53] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 18–33. Springer, 2011.

[54] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[55] Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. Event detection and tracking in social streams. In *Icwsm*, 2009.

[56] Wanita Sherchan, Surya Nepal, and Cecile Paris. A survey of trust in social networks. *ACM Computing Surveys (CSUR)*, 45(4):47, 2013.

[57] Matthew P Simmons, Lada A Adamic, and Eytan Adar. Memes online: Extracted, subtracted, injected, and recollected. *icwsm*, 11:17–21, 2011.

[58] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*, 2017.

[59] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking. *Council of Europe report, DGI (2017)*, 9, 2017.

[60] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2, 2012.

[61] Lilian Weng and Filippo Menczer. Topicality and impact in social media: Diverse messages, focused messengers. *PloS one*, 10(2):e0118410, 2015.

[62] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3:2522, 2013.

[63] Lilian Weng, Jacob Ratkiewicz, Nicola Perra, Bruno Gonçalves, Carlos Castillo, Francesco Bonchi, Rossano Schifanella, Filippo Menczer, and Alessandro Flammini. The role of information diffusion in the evolution of social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 356–364. ACM, 2013.

[64] Zi-Ke Zhang, Chuang Liu, Xiu-Xiu Zhan, Xin Lu, Chu-Xu Zhang, and Yi-Cheng Zhang. Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, 651:1–34, 2016.

[65] Michael Zimmer and Nicholas John Proferes. A topology of twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3):250–261, 2014.

Appendix A

Supplementary Material

In this appendix, we provide in-depth description of the dataset.

A.1 Dataset Collection

Tweets were collected with the Streaming API between September 18th and October 30th of 2017, through the method *filter*. The *’lang’* parameter was set to Portuguese. The *’track’* parameter received the following 137 keywords:

’eleição’, ’eleicao’, ’eleições’, ’eleições2018’, ’eleicoes’, ’eleicoes2018’, ’cirogomes’, ’ciro gomes’, ’cirogomes2018’, ’ciro2018’, ’ciro 2018’, ’lula’, ’lula2018’, ’joãodóriajr’, ’dória’, ’joaodoriajr’, ’joao doria jr’, ’joaotrabalhador’, ’joao trabalhador’, ’doria2018’, ’doria 2018’, ’dória2018’, ’alckmin’, ’marinasilva’, ’marina silva’, ’marina’, ’marina2018’, ’marina 2018’, ’PT’, ’petralha’, ’mortadela’, ’partidodostrabalhadores’, ’partido dos trabalhadores’, ’dilma’, ’dilmarousseff’, ’tucano’, ’PSDB’, ’coxinha’, ’reforma’, ’reformatrabalhista’, ’trabalhista’, ’reformadaprevidencia’, ’previdencia’, ’previdência’, ’rede’, ’redesustentabilidade’, ’rede sustentabilidade’, ’PMDB’, ’DEM’, ’democratas’, ’corrupcao’, ’corrupção’, ’temer’, ’cunha’, ’geddel’, ’policiafederal’, ’policia federal’, ’políciafederal’, ’polícia federal’, ’PF’, ’lavajato’, ’lava-jato’, ’lava jato’, ’sergiomoro’, ’moro’, ’supremotribunalfederal’, ’supremo tribunal federal’, ’stf’, ’gilmarmendes’, ’gilmar mendes’, ’procuradoriageraldarepublica’, ’procuradoria geral da republica’, ’procuradoria geral da república’, ’procurador’, ’pgr’, ’rodrigojanot’, ’janot’, ’raquel dodge’, ’deltandallagnol’, ’ministeriopublico’, ’ministerio publico’, ’ministério público’, ’MP’, ’bolsonaro’, ’bolsomito’, ’bolsonaro2018’, ’jairbolsonaro’, ’MBL’, ’movimentobrasillivre’, ’movimento brasil livre’, ’camaradosdeputados’, ’camara’, ’senado’, ’senadofederal’, ’palaciodoplanalto’, ’desemprego’, ’juros’, ’PIB’, ’produto interno bruto’, ’ministério da fazenda’, ’ministerio da fazenda’, ’impeachment’, ’henrique meirelles’, ’eliseu padilha’, ’moreira franco’, ’blairo maggi’, ’aécio’, ’aecio’, ’politica’, ’política’, ’presidente’, ’presidência’, ’presidencia’, ’executivo’, ’legislativo’, ’judiciario’, ’judiciário’, ’reformapolitica’, ’ministro’, ’partidario’, ’partidário’, ’PEC’, ’votação’,

'votacao', 'crise', 'acordo', 'manobra', 'pedido', 'afastamento', 'denúncia', 'denuncia', 'desvio', 'depoimento', 'divida', 'dívida', 'emenda', 'parlamentar'

We then parsed these data into two CSV files. The first one has one record to each hashtag from each tweet and the second has the mentions network. For example, if a tweet had three hashtags, it was parsed into three different lines in the first CSV file. The lines in this resulting file sums up to 16377202 records, containing 7854621 tweets with 614705 different hashtags, posted from 2348788 users.

A.2 Descriptive Statistics

The frequency of the users and hashtags are distributed as in Table A.1 and A.2.

Table A.1: Distribution of the users' frequency.

N	Mean	Std. Deviation	Median	Minimum	Maximum	Kurtosis
2348788	6.97	98.99	2	1	46702	88859.09

Table A.2: Distribution of the hashtags' frequency.

N	Mean	Std. Deviation	Median	Minimum	Maximum	Kurtosis
614705	26.64	787.06	2	1	413871	135546.6

A.3 Hashtag Co-ocurrence Network

We built a network with the hashtags as nodes, by linking them if they appear together in the same tweet. A total number of 1290755 edges were formed between the 614705 nodes.

With the Python library *networkx* and R package *igraph*, we could run some network metrics. The biggest connected component has 53.3530 % of the graph's size. We observe the degree distribution of this network in Figure A.1. The two following tables show degree and eigenvector [1]centrality measures, respectively:

Table A.3: Top 5 hashtags in degree centrality.

Hashtag	**Degree Centrality**
politica	0.011581
Politica	0.008216
Política	0.007859
PT	0.007198
marina	0.007083

Table A.4: Top 5 hashtags in eigenvector centrality.

Hashtag	Eigenvector Centrality
Rio	0.154544
Brasil	0.141803
Polícia	0.141537
R	0.134673
Região	0.116086

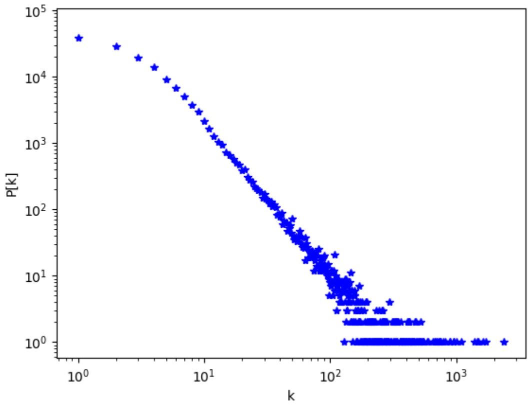


Figure A.1: Hashtag co-ocurrence network degree distribution.

A.4 Mention Network

As stated before, we chose to use the mention network as a proxy to the follower network. Each node is a Twitter user and it is pointing to another node mentioned in any of his/her tweets. Figure A.2 and A.3 are again plots of degree distribution. These figures were generated with the complete network, which have 7882879 nodes and 54745549 edges.

[1] The power method in eigenvector centrality was set to iterate 500 times.

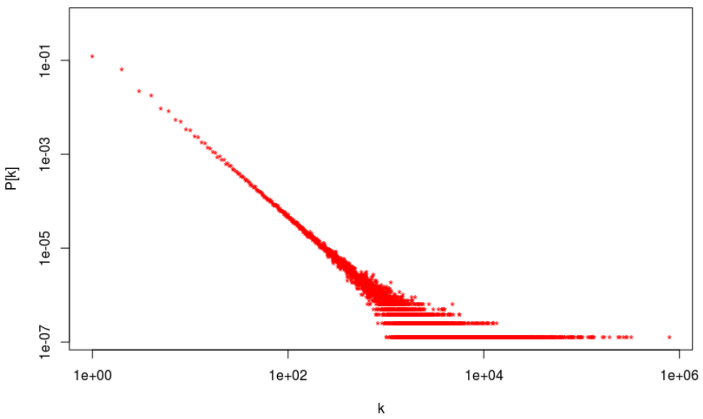


Figure A.2: Mentions network in-degree distribution.

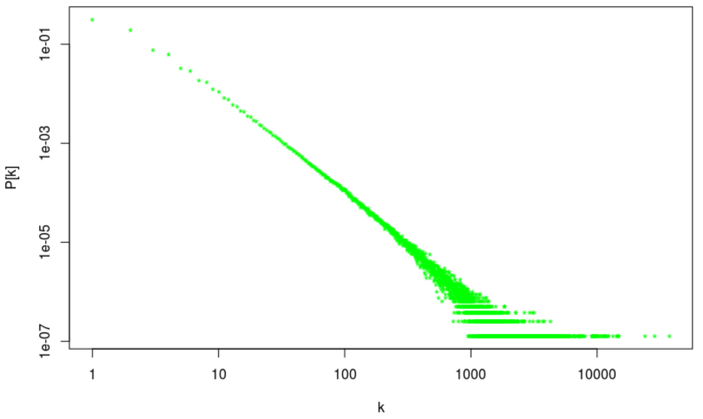


Figure A.3: Mentions network out-degree distribution.

A.5 Filtering

The weight filtering of co-occurrences of hashtags, that is, excluding the edges with one or two joint apperances, returns a network approximately 49% smaller, with 301,207 hashtags. The community detection task was carried out in this filtered co-occurrence hashtag network to access the topics of information.

Users are filtered first for activity. Those who posted with 9 or more hashtags during the entire observation period makes up for 261,521, and not all of them use hashtags from these 301,207, so this number drops to 212,493 users. The next filter for users is to pick those with 3 or more neighbours, because it would not be significant to observe adoptions for those with too few people to adopt from. 157,528 users pass in this filter.

Appendix B

Bias Discussion

In this appendix, we describe what we thought when we obtained the result in Figure 4.1b and the small investigation we carried out to no conclusions. Firstly, we reproduce the same result with a histogram visualization in Figure B.1, and put it against the original distribution of Figure B.2 found in the literature by Cardoso et al [13].

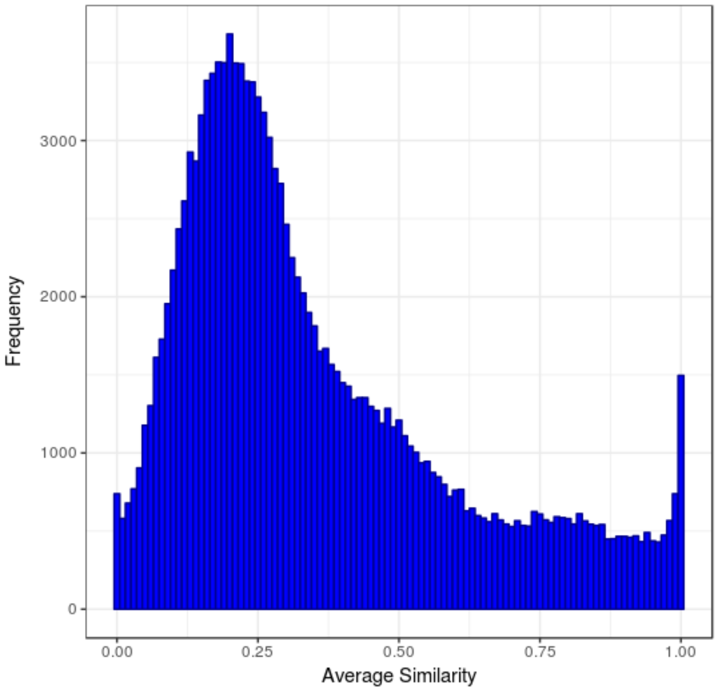


Figure B.1: Number of ocurrences over average similarity, 100 bins.

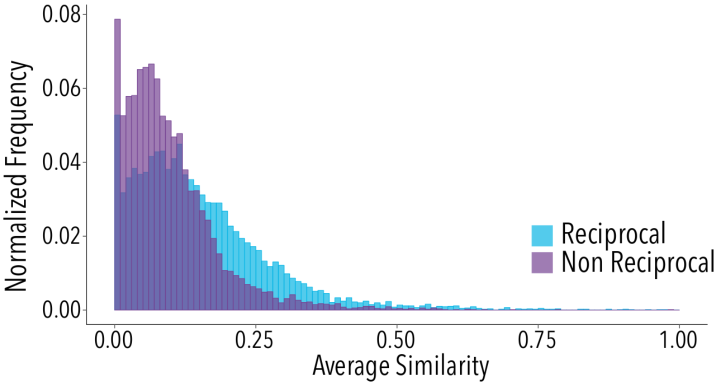
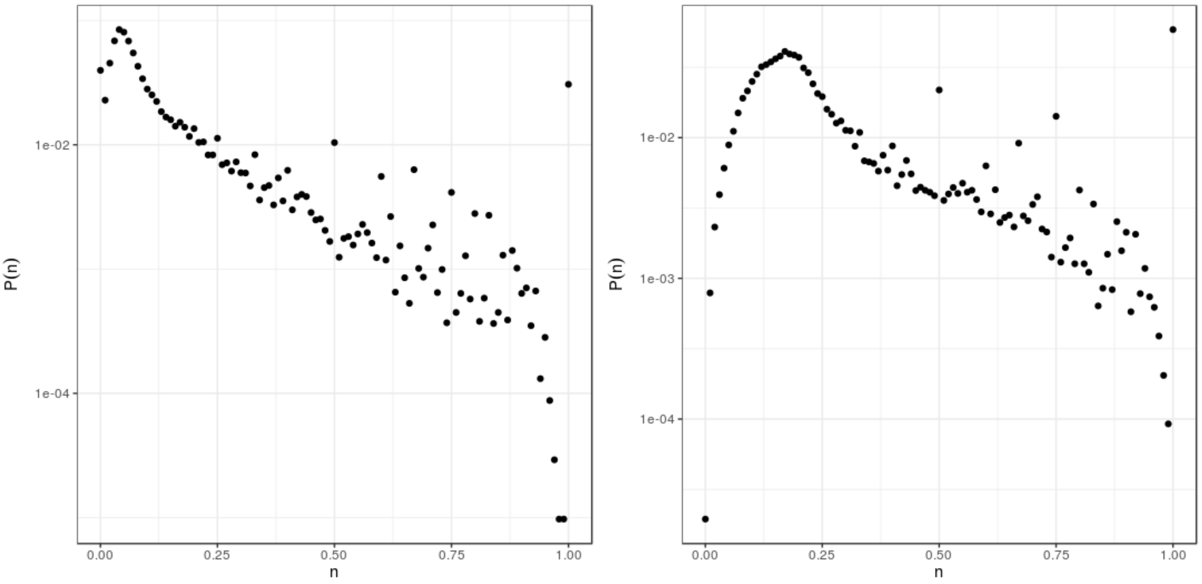


Figure B.2: Reproduced histogram of average similarities in the mentions network. Source:[13]

As it can be seen, we state again the continuous steep increase in our results when the values of similarity averaged by neighbours are past 0.9. We thought this could be cause by Twitter profiles with automatically or semi-automatically controlled behavior dedicated to repost memes of other profiles, namely bots or cyborgs [58]. First of all, we had to ensure they were using the exact same hashtags of their peers and so we looked to the Jaccard index of used hashtags and the topics containing them.



(a) Hashtags

(b) Topics

Figure B.3: Distribution of average Jaccard indices over each user in a semi-log scale.

We can see again a very distinct pattern in the diminishing number of users as the Jaccard index grows and then a sudden spike when the value is 1. A Jaccard index of 1 means that two sets have the exact same elements, and these two sets in this case are the used hashtags in Figure B.3a and the topics associated with them in Figure B.3a, regarding the pairwise comparison of neighbours.

What we did next was to isolate users with 10 or more neighbours and with average similarity bigger than 0.99 to them. These users formed a group of 423 suspect profiles that we decided to check with the Botometer® [1], an online tool created in a joint project by the Indiana University Network Science Institute and the Center for Complex Networks and Systems Research to assign a score of how likely a Twitter account is a bot. Unfortunately, more than half of these users had IDs that no longer existed in the Twitter, and the rest did not yielded significant scores to conclude anything further.

[1] https://botometer.iuni.iu.edu/