UPMC
SORBONNE UNIVERSITÉS

THESIS

submitted in fullment of the requirements for the degree of

DOCTEUR EN SCIENCES

of the

UNIVERSITÉ PIERRE ET MARIE CURIE

spécialité Informatique

Information Diffusion in Complex Networks: Measurement-Based Analysis Applied to Modelling

---

Daniel Faria Bernardes

Defense: 21 March 2014

JURY

Reviewers:	Eric Fleury	Professor, Ecole Normale Supérieure de Lyon
	Marc Tommasi	Professor, Univeristé Lille 3
Examinators:	Sharad Goel	Senior Researcher, Microsoft Research NY
	Bertrand Jouve	Senior Researcher (DR), CNRS
	Pierre Sens	Professor, Université Pierre et Marie Curie
	Emmanuel Viennet	Professor, Université Paris-XIII
Directors:	Matthieu Latapy	Senior Researcher (DR), CNRS
	Fabien Tarissan	Associate Professor, Université Pierre et Marie Curie

Acknowledgments

First and foremost I would like to thank my advisors Matthieu Latapy and Fabien Tarissan, who have assisted me in each step of this journey with incessant enthusiasm and support. They taught me the research trade – *le métier de la recherche* – with great generosity in terms of time and knowledge and for that I am profoundly indebted.

I sincerely thank Eric Fleury and Marc Tommasi for reviewing my dissertation and for Sharad Goel, Bertrand Jouve, Pierre Sens and Emmanuel Viennet for accepting the invitation to integrate the jury of my thesis defense.

I am immensely grateful to the members of the ComplexNetworks team for the warm welcome and stimulating environment in these three years. I thank Clémence Magnien, Jean-Loup Guillaume and Bénédicte Le Grand for the insightful and enjoyable discussions and encouragements. I am thankful for Veronique Varenne's energy and dedication to make practical things happen, in spite of the Byzantine administrative rules. My friends and colleagues Alice, Amélie, Elie, Hamid, Jordan, Lionel, Maximilien, Massoud, Qinna, Raphaël, Romain, Sébastien, Sergey, Thomas and the rest of the LIP6 gang have brightened my time in the University.

Outside the university, I wholeheartedly thank my friends from Polytechnique, specially the Brazilians and the *grimpeurs*, the Neuchâtel gang and my friends and colleagues from CCM and IME/USP, specially Manuel Garcia who inspired me to become a scientist. Valentina, my partner and friend, has been a faithful source of kindness and support, particularly during the final stages of this Ph.D. My deepest gratitude goes to my family for their unconditional love, support and encouragement throughout my life.

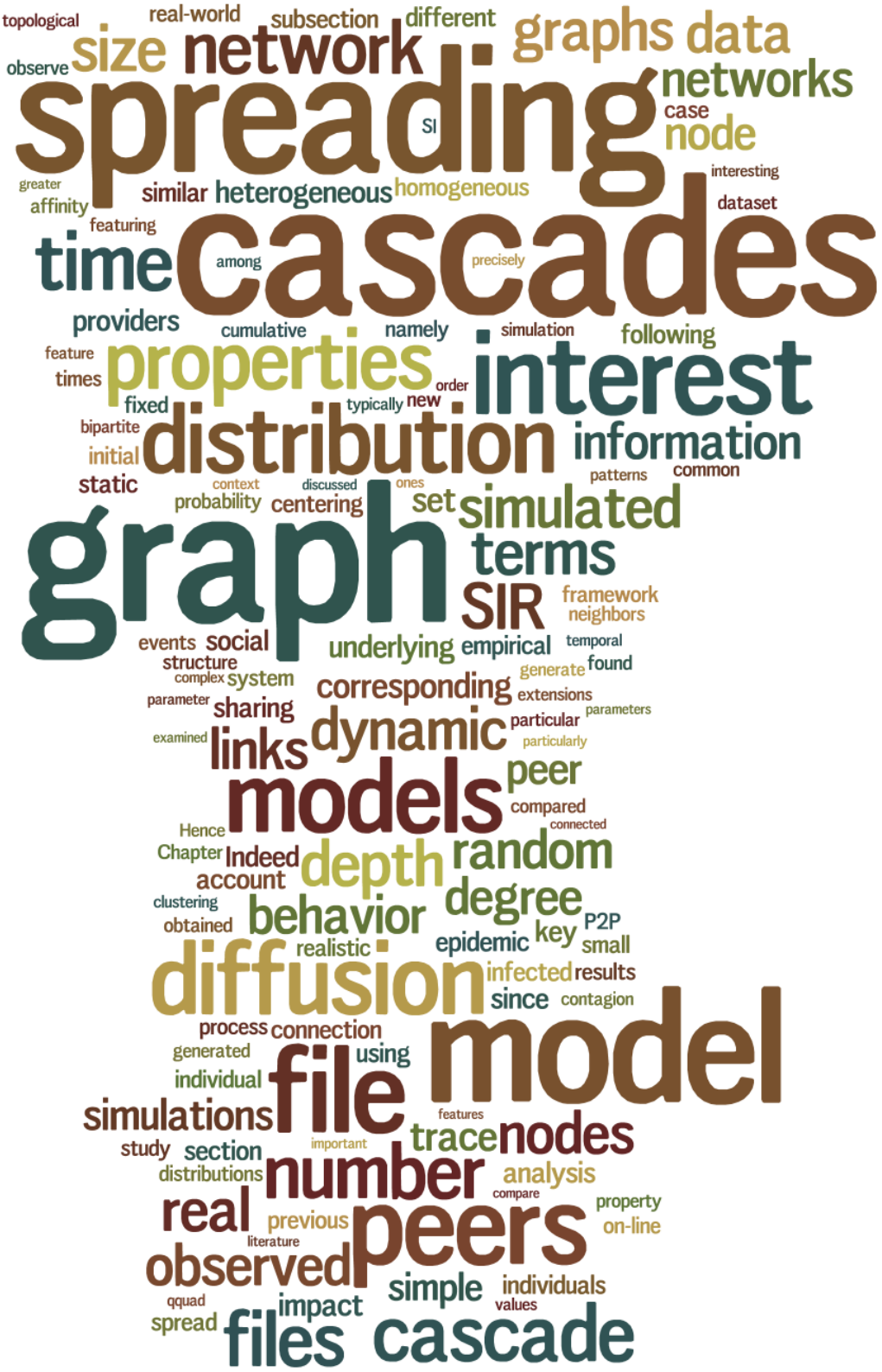


Figure 1 – *Word cloud created with the text of this document.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

On-line social network platforms such as Twitter, Facebook, LinkedIn and many others have become so pervasive today (featuring several hundred million users worldwide) that youngsters may have a hard time imagining the world without them. Even adults may be surprised to realize that the three mentioned platforms were invented less than a decade ago. Indeed, expressions like viral marketing, meme, post, and hashtag have become part of our vocabulary in recent years. Mobile phones and Internet access, which directly and indirectly permeate countless aspects of our daily lives, have also quickly become ubiquitous since their commercialization two decades ago. Simultaneously, the costs of telecommunications and information technologies decreased sharply, allowing real-time information processing and massive tracking of user activity in an effort to create an intelligent and personalized interaction with this omnipresent technology.

User tracking data may be mined by corporations to optimize their operations, to learn customer preferences and offer product recommendations [Leskovec et al., 2007]; it may be used by governments to gather intelligence and monitor crime [Latapy et al., 2013] and by the general public to obtain detailed information on disasters and riots [Ball, 2011, Doan et al., 2012]. This increasing digital presence and information sharing also raises novel issues regarding the relationship of individuals and their data, notably in terms of intellectual property law and privacy [Lessig, 2002, Fertik and Thompson, 2010]. In addition to the highly valuable

practical implications mentioned, massive user tracking offers a unique opportunity to study large-scale networks and emerging complex collective behavior from local interactions therein.

The interest in emergent behavior and complex pattern formation has a long history in philosophy and natural sciences, being discussed notably by Aristotle [Tredennick, 1933], John Stewart Mill [Mill, 1843], the economist Frederick Hayek [Hayek, 1948] and the evolutionary biologist Julian Huxley [Huxley and Huxley, 1947] among other renowned thinkers. The challenge of dealing with complexity is inherent in scientific inquiry, although conventional scientific fields focus on particular "scales of reality". For example, although in theory animal behavior could be studied in terms of atom interactions – since organisms are composed of cells, which can be described in terms of molecule interactions, which are ultimately made up of atoms – in practice the study of each of these scales is done by specialized fields, namely ethology, cell biology, biochemistry and molecular physics. Nonetheless, as the applied mathematician-turned-social scientist Duncan Watts points out:

"Increasingly, the questions scientists find more interesting – from the genomics revolution to the preservation of ecosystems to cascading failures in power grids – are forcing them to consider more than one scale at the time, and so to confront the problem of emergence head-on" [Watts, 2011].

This micro-macro issue is not restricted to the natural sciences: it is fundamental to economics [Smith, 1789, Schumpeter, 1909, Klein, 2012] and social sciences [Granovetter, 1978, Ritzer, 2007]. Indeed, individuals are embedded in social networks and interactions occur in within this network. However, characterizing and analyzing social ties in detail has been historically laborious and generally impractical to implement at large-scale. Hence the possibility to leverage the information technology and the on-line social interaction data to shed light on micro-macro questions has interested a growing number of researchers. Interestingly, these researchers are not only social scientists, but also mathematicians, computer scientists, physicists and others. See [Freeman, 2004] for a comprehen-

sive account of social network analysis development.

Among the various instances of micro-macro issues in the intersection of natural and social sciences is the phenomenon of diffusion and cascading behavior: this phenomenon is characterized by the spread of information through a process of individual-to-individual contagion. Though contagion and social influence were a concern since ancient times, the systematic analysis of these phenomena was ignited in the late 19th and early 20th century with contributions from social sciences [Le Bon, 1895, de Tarde and Parsons, 1903] and epidemiology [Kermack and McKendrick, 1927]. Since then, new models and refinements appeared along with new empirical data. However, until recently sufficiently detailed large-scale data was unavailable to validate the micro foundations of diffusion models.

Diffusion phenomena are a class of propagation phenomena characterized by the spread of information or physical objects through some process of individual-to-individual contagion: classics examples are biological viruses such as HIV. In this case, starting with a small number of infected individuals, this virus spread to thousands of individuals in a relatively small amount of time. This bursty behavior is common to many viruses, but it is also possible that diffusion spreads slowly or that it dies out before reaching a significant portion of the population. See example in Figure 1.1.

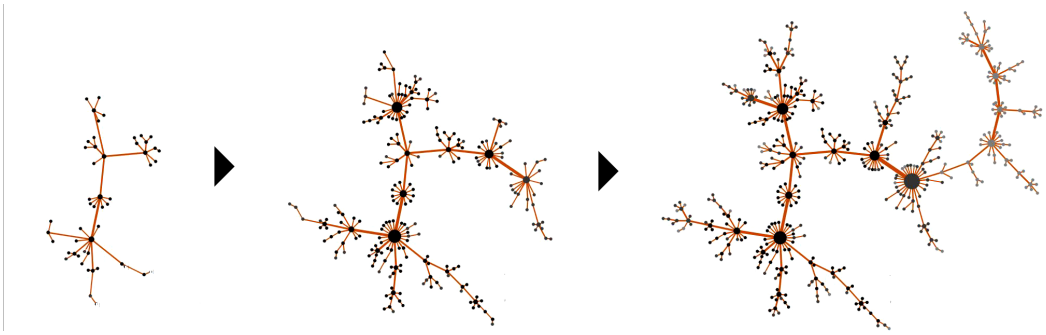


Figure 1.1 – *A spreading cascade example: early, middle and final stages. This was an on-line experiment where an applet called "Happy flu" spread among users. [Friggeri et al., 2011].*

This thesis consists of a data-driven investigation of a real-world information diffusion on a large-scale social network using on-line file sharing traces. More

precisely, we identify key real spreading cascade properties and examine the capability of standard models to reproduce these properties.

In the following, we present a survey on diffusion and cascading behavior studies and we end with an overview of the thesis and its contributions.

1.1 Context and Survey

1.1.1 Complex networks

A common feature in the micro-macro issues mentioned in the previous section is the presence of pairwise relations between parts of a system, which can be modeled in terms of networks. Examples include food webs and ecological networks, power grids and the World Wide Web, friendship and collaboration networks, which are instances of biological, technological and social networks. These networks are rich objects in their own right and often constitute the structure on which interesting dynamic takes place, such as "viral" diffusion and content search. More formally, we can represent these networks with the mathematical notion of graph, denoted $G = (V, E)$ and characterized by a set of *nodes* V and a set of *links* between nodes E, which can be directed or undirected (undirected links are also called *edges*). Although the term "network" may occasionally imply more information than "graph" in certain scientific communities, in this work we consider both terms as synonyms and use them interchangeably. See [Bollobás, 1998, Diestel, 2010] for modern references on graph theory and [Easley and Kleinberg, 2010] for applications.

Real-world networks

In recent years a growing number of empirical studies of real-world large-scale networks have been developed, particularly taking advantage of on-line platforms. Wikipedia is a case in point: this open, collaborative on-line encyclopedia has tracked the activity of each editor and each entry since its inception. This allowed

the study of the collaboration network of editors, connected if they worked on the same entry [Crandall et al., 2008, Kittur and Kraut, 2008]. On-line games also provide novel instances of such networks, e.g., the graph of World of Warcraft players who have taken part in common raids or activities [Wotal et al., 2006]. Scientific collaboration has attracted attention before the Internet, notably with the works of Derek de Solla Price – who developed a theory of the growth of citation networks, based on what would now be called a preferential attachment process [Price, 1976]. This theme was echoed and expanded with the advent of on-line indexing platforms [Newman, 2001].

Another class of real-world networks, similar to citation networks and loosely termed information linkage graphs are characterized by massive and diverse datasets, typically from the World Wide Web: nodes are pieces of information linked together. Examples include the network of web pages connected by hyperlinks [Kleinberg et al., 1999, Huberman and Adamic, 1999], blogs and linkages among bloggers [Kumar et al., 2004, Leskovec et al., 2007, Salah Brahim et al., 2011], product reviews and users on shopping sites [Guha et al., 2004] and Twitter accounts and followers [Sharma et al., 2012]. An important example of networks which reveal social ties and infrastructure are communication networks, which represent individuals who have had a recorded conversation. Evidently, the content of the conversations is generally private and therefore inaccessible or hidden on purpose to preserve anonymity; instead, researchers generally work with metadata concerning these interactions, particularly who-contacted-whom, occasionally associated with a time stamp. Empirical studies of these networks include long-range communications, such as mobile phone [Onnela et al., 2007], students exchanging emails [Kossinets and Watts, 2006] and instant-messaging [Leskovec and Horvitz, 2008] as well as short range communications, such as the contact networks of participants in a conference [Isella et al., 2011] and in a roller skate event [Neiger et al., 2012].

We close this section with an important class of real-world networks: natural networks, particularly from biology. Examples include the structure of neural connections within an organism's brain [Sporns et al., 2004], food webs – nodes represent species and links prey-predator relations [Dunne, 2006] – and metabolic

networks – nodes are functional compounds and links chemical interaction between them [Barabási and Oltvai, 2004].

Structural properties

Despite the generality of the characterization of networks and the diversity of contexts they are found, the empirical works mentioned previously found a set of non-trivial structural (or topological) properties common to a wide-range of observed real-world graphs, mainly: small diameter, heavy-tailed node degree distribution and global sparsity/local density. Thus, these networks were generally labeled *complex networks*.

A graph is said to feature a small diameter if for each pair of nodes there exists a path connecting with whose length smaller than some small constant. In the context of social networks this property has a relatively long story, beginning with the 1929 play *Chains* by Hungarian author Frigyes Karinthy, where the concept appeared stylized as "six degrees of separation,": i.e., that any two individuals could be connected through at most five acquaintances generally. In academia, this concept was echoed in a landmark experiment by Stanley Milgram in 1967: he sent several packages to 160 random people living in Omaha, Nebraska, asking them to forward the package to an acquaintance who they thought would bring the package closer to a final individual in Boston, Massachusetts. He reported that chains varied in length from two to ten intermediate acquaintances, with a median of five intermediate acquaintances [Travers and Milgram, 1969]. A recent study reported a similar value: the average chain of contacts between users of Microsoft instant-messaging system was 6.6 people [Leskovec and Horvitz, 2008]. The same property was observed in citation networks [Newman, 2001], the collaboration network of actors [Watts and Strogatz, 1998] and elsewhere [Kleinberg, 2006].

In terms of node connectivity, recent studies have observed heavy-tailed [1] *node degree distribution*, i.e., the frequency of the number of node neighbors in

1. Heavy-tailed distributions are probability distributions whose tails are heavier than the exponential distribution. More precisely, let X be a random variable with distribution F on $\mathbb{R}$ and tail function $\bar{F}(x) = \mathbf{P}(X > x)$, $x \in \mathbb{R}$. The distribution F is *heavy-tailed* if $\limsup_{x\to\infty} \bar{F}(x)e^{\lambda x} = \infty$ for all $\lambda > 0$. See [Foss et al., 2013] for further properties.

the graph [Newman, 2010]. In particular, some studies have reported power-law degree distributions [2] in real-world graphs such as the autonomous system of the internet [Faloutsos et al., 1999] and the Web [Kleinberg et al., 1999, Barabási and Albert, 1999, Adamic and Huberman, 2001]. Power-law distributions – sometimes referred to as "scale-free" distributions – had been found in different contexts, notably counting the frequency of in natural languages [Zipf, 1948] and examining income distributions [Pareto, 1897]. Despite universal character, a number of studies in network analysis have been questioning the empirical methods used to fit such distribution [Clauset et al., 2009, Kolaczyk, 2009]. In particular, Jackson and Rogers show how some allegedly scale-free degree distributions are better fitted by other heavy-tailed distributions [Jackson and Rogers, 2005]. In sum, though heavy-tailed degree distributions have been observed consistently, well fitted power-law distributions have been shown to be rarer.

Complex networks have also been reported to be globally sparse – meaning that nodes are typically connected to few other nodes –, but featuring high local density, measured in terms of a *clustering coefficient* [Watts and Strogatz, 1998] [3]. More precisely, these networks feature a high clustering coefficient relative to what would emerge if links were determined by an independent random process [Newman, 2001]. Ideas behind clustering have been important in social sciences since Simmel [Gurcel and Watier, 2002], who pointed out the interest in triads (triples of multiple connected nodes). Empirical results have found high local clustering in actor collaboration networks [Watts and Strogatz, 1998], in the Web [Adamic, 1999], in dating networks [Liljeros et al., 2001] and other places. In connection to this property is the question of community detection, which spawned an entire field of research dedicated to develop methods to cluster nodes in terms of their connection patterns (See [Fortunato, 2010] for a comprehensive account of the field).

2. A (positive) power law distribution is a heavy-tailed distribution featuring a tail function which is asymptotically given by a power-law, that is: $\bar{F}(x) \sim (x_{\min}/x)^{\alpha}$ as $x \to \infty$, with a scale parameter $x_{\min} > 0$ and a shape parameter $\alpha > 0$. It has all moments of order $\gamma < \alpha$ finite, while all moments of order $\gamma \geq \alpha$ are infinite.

3. The local clustering coefficient C_i for a vertex $v_i \in V$ is then given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. Let N_i is the set of neighbors of v_i and $d_i = |N_i|$ its degree. The corresponding local clustering is $C_i = |\{(v_j, v_k) \in E : v_j, v_k \in N_i\}|/(d_i(d_i - 1))$.

1.1.2 Diffusion and cascading behavior

Researches have been trying to characterize and model epidemic dynamics systematically since the early 20th century. Other individual-to-individual spreading phenomena have attracted the attention of researchers, namely the spread of ideas and social norms. More recently, a number of works reported the diffusion of on-line information such as links, files and memes. Asserting the influence of individuals and the detailed mechanisms of contagion in such contexts is challenging. The first issue is to decompose the evolution as a result of multiple individual interactions. Secondly, even when we can model the spread in terms of individual actions, it is not obvious how to model individual behavior. Nevertheless, researchers developed models inspired in epidemiology to describe general features of social diffusion phenomena.

At the same time, empirical studies have documented a wide range of diffusion phenomena with increasingly more detail. Indeed, the punctual spread of information from individual to individual has been difficult to observe until recently. This picture has changed with the emergence of on-line platforms: one may observe the spreading in greater detail. In the following, we will present a survey of theoretical and empirical results on diffusion and the main open challenges in the domain.

Diffusion models

The first diffusion model formalized in mathematical terms, which captures the spreading as described above, was proposed by Kermack and McKendrick [Kermack and McKendrick, 1927, Anderson and May, 1991, Andersson and Britton, 2000], focusing on the global evolution of the infected population. In its simpler setting, the model partitioned the population in two groups, a susceptible and an infected group, and made a few assumptions on the spreading behavior, namely that the number of infected individuals growth is initially proportional to the current number of infected individuals, reaching saturation point as most individuals become infected. These relations were formalized in terms of deterministic differential equations which could be used to determine the long term behavior of

the disease. Extensions of this model considered supplementary population partitions corresponding to other classes of individuals such as exposed or recovered individuals (or removed individuals if the disease is lethal). This compartmental models are globally known in the literature as compartmental models or simply SIR models, in reference to the widest-known model in this category.

Outside the context of epidemiology, a landmark model was proposed by Bass in 1969 [Bass, 1969] for the adoption of innovations. The analogy with epidemic models is explicit: an infectious object (an idea, a product or a behavior) is assumed to spread from infected to susceptible. The goal of this model is also to capture the global dynamic of the population in terms of these classes of individuals, similarly to the SIR models. The evolution of one compartment depends on the relative size of the other compartments, and it is formalized with differential equations. The key feature of this model is the adoption curve, a S-shaped curve which tracks the fraction of adopters relative to the total population over time.

Both models were formulated with limited support from empirical observations and thus assumed a quite general spreading behavior. Indeed, an underlined hypothesis of both models is that individuals in one compartment have the same influence on another compartment, that is, influence is distributed uniformly in the population. Social network analysis, however, has shown that certain individuals are much more connected than others. In addition, we know that the behavior of individuals can be quite heterogeneous: for example, in the context of sexually transmitted diseases, some individuals have much more partners and are more active than others. To account for these heterogeneities, in recent years these models were adapted to feature the underline network structure of individuals: in this case, the spreading behavior is centered on the individual and his or her neighborhood in the network.

Epidemic diffusion was adapted to the network setting, drawing from percolation theory: in this case, it is assumed that an individual can only infect its neighbors on the network. With these models it is possible to describe the same quantities as those of previous models, namely the fraction of infected individuals, aggregating the local behavior of every individual of the population. Since the

adaptation of these models to this context, a number of studies was performed investigating the global asymptotic behavior of the epidemic in terms of the network topology. In particular, this was simulated on a number of real world networks. Novel questions were addressed such as, given a certain network, what is the optimum vaccination strategy or how to select the best set of initial nodes to ignite an epidemic [Pastor-Satorras and Vespignani, 2001, Kempe and Kleinberg, 2002, Leskovec et al., 2007].

An alternative model based on local dynamics was introduced in 1978 by Granovetter [Granovetter, 1978] and improved upon recently [Kempe and Kleinberg, 2002, Dodds and Watts, 2005]. Like the network version of the SIR models, Watts and Dodds' model also makes an assumption on the spreading behavior of individuals. However, instead of assuming that one individual may infect its neighbors, this model assumes that each individual adopts a piece of information if a certain number (or fraction) of its neighbors adopted it as well – i.e., an individual's decision is triggered by its surrounding. These adoption/threshold models also yield cascading behavior, but they are not equivalent to epidemic models [Dodds and Watts, 2005]).

Empirical studies

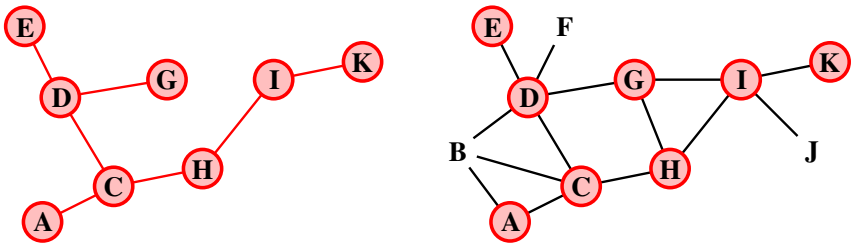
Until recently, empirical data on diffusion phenomena, mainly epidemic outbursts and product adoption, consisted of aggregated data like the number of infected individuals on a given time. As mentioned previously, identifying punctual individual to individual transmissions is challenging at large scale. In some cases, like for the influence exerted by individuals, this is not directly accessible, so empirical studies have focused on proxy measurements and assumptions about the link between these measurements and influence itself.

Recent technology, particularly on-line platforms, allowed the observation of detailed large-scaled diffusion phenomena, namely diffusion on blogs [Adar and Adamic, 2005], e-commerce [Leskovec et al., 2006], mobile phone networks [Onnela et al., 2007], on-line gaming [Bakshy et al., 2009] and social networking [Sun et al.,

2009, Bakshy et al., 2011, Goel et al., 2012]. These new datasets also uncovered the diffusion trail that is not only the information of who received an information at a given time, but also by whom this information is sent. Hence, for a given spreading information, one can construct the corresponding spreading cascade, a directed graph connecting infected individuals with whom they obtained the information. In particular one can compute the length of the path connecting the original source of information to any given node in this graph and related measurements such as the maximum path length between two nodes of the graph, termed the cascade depth, and the internal density of the cascade.

Another important aspect of these new data is the possibility to compare and improve diffusion models. As mentioned previously, these models have local spreading rules, which depend upon the underlying network. Therefore, in order to test the local spreading assumptions, it is necessary to know both the underlying network and the spreading trace. Until recently, there have been few examples in the literature of open data sets of large scale diffusion phenomena featuring the diffusion trace and the underlying network. In some cases you may know the network but miss the complete diffusion trace. An example is the diffusion of content in blogs: in a highly heterogeneous media environment, content diffusion is likely a combination of interpersonal spreading and more traditional media channels. In other words, people may post something on a blog after seeing it on a friend's blog or after seeing it on television or somewhere else. In other cases, on the contrary, you may miss the network, but know the entire diffusion trace – e.g. e-mail spreading [Liben-Nowell and Kleinberg, 2008] and the Happy flu experiment [Friggeri et al., 2011]. See Figure 1.2 for a schematic illustration of empirical shortcomings observing spreading cascades.

Parallel to the advent of newer and more detailed (albeit not generally open) datasets featuring the spreading cascade and the underlying network, a novel line of research focused in the reconstruction of the underlying network, using the spreading cascade and maximum of likelihood optimization techniques [Gomez-Rodriguez et al., 2012]. Important research effort has also been dedicated to the characterization of on-line diffusion in terms of spreading cascades: recently Goel et al. proposed different structural metrics to differentiate diffusion spreading to



(a) *Spreading cascade but no underlying network.* **(b)** *Nodes reached by diffusion but no transmission links.*

Figure 1.2 – *Spreading cascades empirical shortcomings.*

media broadcast in on-line social networks [Goel et al., 2013].

Finally, despite the variety of diffusion models available, considerably little attention has been devoted to the development of estimation methods to calibrate those models. These techniques are fundamental for applications and to assess the pertinence of the models, given empirical datasets. In this sense, we have identified two main papers on the subject: one proposing a maximum likelihood techniques to estimate SIR and adoption models [Saito et al., 2008], but whose framework is costly and not scalable and [Goyal et al., 2010] which propose an interesting framework for epidemic models.

1.2 Summary and contributions

As discussed in the literature survey, the understanding of diffusion phenomena has undergo major improvements since its beginnings. A key improvement came with the introduction of network analysis, which integrated social and technological networks with the spreading process. More recently, with the advent of large-scale on-line platforms which keep track of user activity in great detail, the empirical focus has gradually been shifting from simple aggregated statistics, such as number of infected individuals, to more complex objects, such as spreading cascades, which encode the diffusion trail. Improvements in diffusion models followed, but most theoretical results rely on asymptotic analysis and equilibrium / steady state conditions. Given the complexity of diffusion phenomena, challenges exist

in numerous fronts: identify the most relevant structural metrics for spreading cascades, determine the individual influence and spreading behavior, establish compatible models capable of producing realistic spreading cascades.

This thesis presents the following contributions in this context. In Chapter 2, we identify and obtain a large-scale diffusion trace with a detailed information of who transmitted the information to whom: file sharing logs in peer-to-peer (P2P) network. This level of information is key to assess the hypothesis of standard diffusion models. Another crucial information is the underlying network where the diffusion takes place: to this end, we present a framework to reconstruct the social network of users in this system, related by common interests. We compute structural statistics for this network and report the same properties featured by typical complex networks, as discussed in the previous section.

In Chapter 3, we analyze the most standard diffusion model in the literature and in the context of P2P networks, the SIR model. Supposing the observed spreading cascades were essentially generated by a process with the dynamic of this epidemic model, we calibrate the model parameters with the data, perform model simulations and compare them to the real cascades. We show that this model is unable to reproduce key topological features of spreading cascades. Moreover, this observation remains true for natural extensions of this model, featuring peer and file heterogeneities. We also propose an affinity measure to refine the underlying network and analyze spreading cascades in this refined graph.

In Chapter 4 we demonstrate the importance of taking into account temporal patterns both in terms of underlying network and of the spreading process. We show how the dynamic interest graph can be reconstructed from the original interest graph and the connection pattern from users and that it is a key ingredient to generate realistic cascades in terms of size.

In addition to this empirical study, in Chapter 5 we analyze the impact of the underlying network structure on simulated spreading cascades using the models discussed in the previous chapters. In the literature, a substantial amount of interest was given to the asymptotic behavior of the number of infected individuals.

We examine this question from complementary perspective, investigating the evolution of the cascade structure in a constrained in time, as the ones observed in our dataset. In addition, instead of focusing exclusively on the number of infected individuals, we investigate the cascade structure in terms of the three cascade properties discussed in the previous chapters. In sum, we assessed the impact of key topological properties in time-bounded contagion spreading and observed that the distribution of the number of neighbors of seed nodes had the most impact in our setting.

We conclude in Chapter 6, summarizing the results obtained and discussion the perspectives opened by this study. In particular, we explore new avenues in empirical analysis of spreading cascades, improvements to the framework used and general questions related to the study of information diffusion.

Chapter 2

Dataset and Framework

Contents

In recent years on-line platforms have registered a vast amount of detailed interaction data. This rich data enticed scientists interested in information diffusion to better characterize large-scale diffusion and examine the long held assumptions and models on the subject. We subscribe to this move, studying the diffusion of files in a peer-to-peer (P2P) file sharing system. In this chapter we present the dataset used throughout this thesis, describe how it was obtained and the framework to reconstruct from it the spreading trail and the underlying network.

2.1 Measuring real-world information diffusion

As we have discussed in the previous chapter, standard contagion models are based on local transmission rules which depend upon the structure of the

underlying network, so in order to study it empirically the data must features both the spreading trail (who spread what to whom at what time) and the underlying network. Since the beginning of this thesis, a number of datasets meeting this criteria appeared in the literature, particularly in the context of on-line social networks [Bakshy et al., 2011, Dow et al., 2013, Goel et al., 2013]. Some rich datasets existed previously, but were typically proprietary [Leskovec and Horvitz, 2008]. However, at the beginning of this thesis, a lot of attention was given to the study of information diffusion on the web, particularly the citation links in blogs, which were publicly accessible to anyone in the scientific community. However, reconstructing the diffusion trail from citation links has its shortcomings as the following example illustrates: a blogger views a video link on blog X and posts it on his blog Y with reference to blog X; another blogger sees the post on blog B and decides to post the video link on his blog Z with reference to the original post on blog X. That is, the blog Y was "shortcut" in the observed spreading trail, giving the impression that the author of blog Z obtained the information directly form X, when in fact the information spread through Y. This measurement issue undermines the empirical analysis of information diffusion, so to overcome it we decided to study diffusion on a peer-to-peer file sharing systems, setting up a novel large-scale diffusion dataset which we make publicly available on-line [1].

2.1.1 Peer-to-peer file sharing systems

Peer-to-peer file sharing systems have evolved into a large traffic source in the Internet and established themselves as an important platform for content distribution [Sen and Wang, 2004, Ban et al., 2011]. They constitute a remarkable case of interaction between a technological layer (network of computers) where the traffic occurs, and a social layer (overlay network of peers, structured by related interests) where the content spreading occurs. In eDonkey file sharing systems, one of the main P2P file sharing systems, peers connect to a server to query for files of other connected peers and to provide files to fellow peers upon request [Kulbak et al., 2005]. More precisely, file sharing can be divided in three steps, which we denote, respectively, *textual query*, *file request* and *P2P file exchange*. First, the client makes a textual query to the server, which returns a list of available files in the system

1. Dataset available at: http://www-complexnetworks.lip6.fr/~bernardes/p2pdata2d

(each represented by a unique hash code) whose description matches the textual query. Next, the peer will choose a subset of files in this list and make a second query to the server requesting the unique id of potential providers for each selected file. Finally, the client contacts the providers directly and transmission between them ensues.

In this system, the first two steps described above can be observed at the eDonkey server level, as all file requests are intermediated by the server. Evidently, the file exchange step itself cannot, as the communication is done peer-to-peer. Nonetheless, it is possible to track the file diffusion in the system monitoring the corresponding requests preceding each P2P file exchange since users and files are uniquely identified by the server. Each file request is decomposed in individual events which are encoded as 4-tuples in the following format: (t, P, C, F), where capital letters represent unique ids. Such a tuple accounts for a request made at time t of the file F by the peer C, satisfied by the peer P. In other words, P is a provider of the file F pointed out by the server to the peer C at time t. The spreading trace is composed of all these individual events, as the example in Table 2.1 illustrates.

Time	Provider	Client	File	Time	Provider	Client	File
1	1	2	A	8	5	8	A
2	1	3	A	9	4	9	D
3	4	2	B	10	4	3	B
4	4	1	C	11	10	5	E
4	5	1	C	12	9	11	F
5	1	4	D	13	4	9	C
6	5	6	A	14	4	12	D
6	3	6	A	15	9	7	F
7	1	7	C				

Table 2.1 – *Example of spreading trace featuring 12 peers and 6 files. The trace is comprised of 17 events, displayed in chronological order. Each event represented by a 4-tuple composed of a timestamp in seconds, two peers (a provider and a client) and a file. Files are represented by letters and peers are numbered from 1 to 12. Any peer can be a client, but only peers who possess a file can be its provider (either the peer possessed it before entering the P2P system or acquired it by sharing in the system).*

2.1.2 Measurement and analysis

We have obtained a diffusion trace recording these events at the eDonkey server level, akin to [Aidouni et al., 2009], anonymized due to privacy concerns. We have parsed the raw measurements in XML and filtered to the format described previously. Monitoring a contiguous time window of $T = 170353$ seconds (approximately 48 hours) we have observed $5\,380\,616$ peers, $1\,986\,588$ files and $471\,411\,593$ file request events. The requests (represented by the tuples described previously) can be grouped in terms of peers or files, as illustrated in Figure 2.1, revealing temporal patterns which will be explored in further detail in Chapter 4.

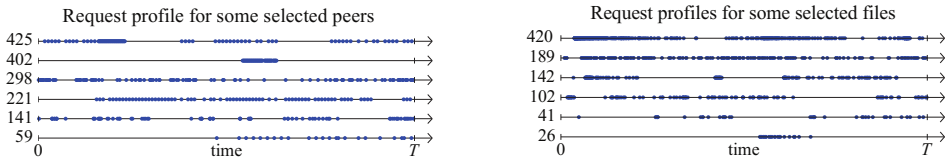


Figure 2.1 – *Request profiles of six clients and six files. Dots represent requests in each time line corresponding to a peer or a file.*

The estimation methods and simulations proposed in the following chapters are numerically expensive in terms of resources, so we have decided to work with a subsample of this dataset, corresponding to the first 8 hours of measurements, which is still large-scale in terms of the number of peers, files and transmission events, but that could be manageable without an enormous engineering infrastructure. Indeed, let $\mathcal{P}$ be the set of all peers and $\mathcal{F}$ the set of all files exchanged in our subsample. We have $|\mathcal{P}| = 1\,908\,500$ peers, $|\mathcal{F}| = 801\,280$ files and $22\,944\,800$ file transfer events.

The Figure 2.2 shows, events (individual file requests) arise almost linearly with time and the number of registered peers and files follows a trend with fluctuations which may be due to circadian cycle patterns. Thus, the subsample in question preserves important characteristics of the original dataset in terms of rate of observations of new events, peers and files. Circadian patterns, however, are likely unobservable in the subsample, but the shorter time window may offer counterbalancing advantages. Namely, it reinforces the likelihood that essentially all the spreading of files in the period were due to file sharing on the network. Indeed, peers receive and share files not only through P2P file sharing systems but

they also do it through other non-observable means, such as using physical devices. However, file sharing in different channels is not done in the same speed nor with the same frequency, and though we cannot guarantee that there was no interference due to off-line sharing, it seems reasonable to neglect it in this time window.

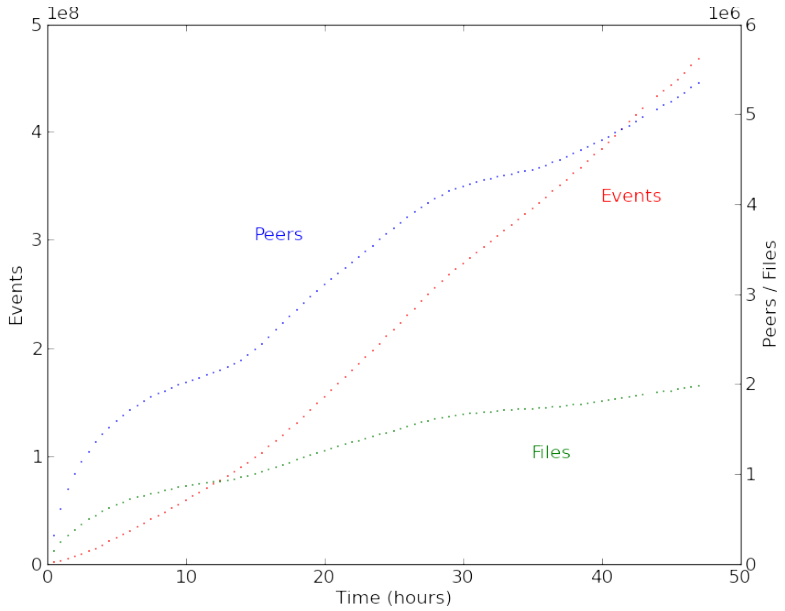


Figure 2.2 – *Evolution in the number of observed events (individual file requests), peers and files in the P2P sharing system during 48 hours of contiguous measurement.*

Let $\mathbf{D}$ be the set of all recorded tuples in the subsample (henceforth denoted simply *dataset*). Before we begin a more structured study of diffusion in the next section we highlight some basic file sharing statistics of the trace $\mathbf{D}$. First, we present two statistics related to the typical number of interested peers per file: the median number of interested peers per file, 5, and the average number of interested peers per file, 14.73, with standard deviation 34.74. Second, we estimate the number of files commonly shared by peers: median number of files shared by peers is 3 and the average is 6.19, with corresponding standard deviation 12.66. These values suggest an heterogeneous distribution for both properties, as we shall see later in the following sections. Another important aspect of our P2P trace in terms of file sharing statistics is the abundance of *free-riders* – that is, peers who benefit of shared files in the system, but who do not share back. In our dataset, while 99.63% of the peers are clients (i.e., have requested a least one file) only 4.33% of them

have supplied files.

Heterogeneous file sharing behavior and high proportion of free-riders have been observed in the literature, in P2P file sharing systems. A measurement study of the Gnutella file sharing system [Adar and Huberman, 2000] found that approximately 70% of peers provide no files and that the top 1% of the peers provide approximately 37% of the total files shared. Similar patterns have been observed in subsequent studies of Napster and Gnutella system [Saroiu et al., 2002]. In 2005, [Hughes et al., 2005] found free-riders have increased to 85% of all Gnutella users. Similar patterns were also observed in the eDonkey system [Handurukande et al., 2006].

2.2 Spreading trace

The focus of this work is the study of real-world diffusion, in terms of its spatiotemporal structure. In oder to make this notion precise, be begin defining the main object of analysis, namely the *spreading cascade*, which represents the diffusion trail of each file in the P2P system, as recorded in the spreading trace. We also identify the *initial providers* or *seeds* for each file, which will be necessary in later chapters.

2.2.1 Spreading cascades

For a file F, the spreading cascade is a directed graph featuring the set $\mathcal{P}_F$ of peers who have participated in the spread of F (as clients and/or providers) and links $P \to C$, connecting each client C with the first peer(s) who provided F to it. More formally, let $\tau_F(C) = \inf\{t : (t, \cdot, C, F) \in \mathbf{D}\}$ be the first instant C obtained F and let the directed graph $\mathcal{K}_F = (\mathcal{P}_F, \mathcal{L}_F)$ be the spreading cascade of F, with

$$\mathcal{P}_F = \{P \in \mathcal{P} : (\cdot, P, \cdot, F) \in \mathbf{D} \vee (\cdot, \cdot, P, F) \in \mathbf{D}\}$$

$$\mathcal{L}_F = \cup_{C \in \mathcal{P}_F}\{(P, C) \in \mathcal{P}_F \times \mathcal{P}_F : (\tau_F(C), P, C, F) \in \mathbf{D}\}$$

A client requesting a file may receive a response from potentially several providers simultaneously, which implies that nodes in the cascade graph not only have multiple outgoing links, but also multiple incoming links in general The causality induced by the fact that we only consider the links corresponding to the first time a node received F prevents the appearance of cycles. Hence the cascade is in fact a directed acyclic graph (DAG). As an example, in Figure 2.3 we construct a spreading cascade for each file in the spreading trace in Table 2.1.

Time	Provider	Client	File
1	1	2	A
2	1	3	A
6	5	6	A
6	3	6	A
8	5	8	A
3	4	2	B
10	4	3	B
4	4	1	C
4	5	1	C
7	1	7	C
13	4	9	C
5	1	4	D
9	4	9	D
14	4	12	D
11	10	5	E
12	9	11	F
15	9	7	F

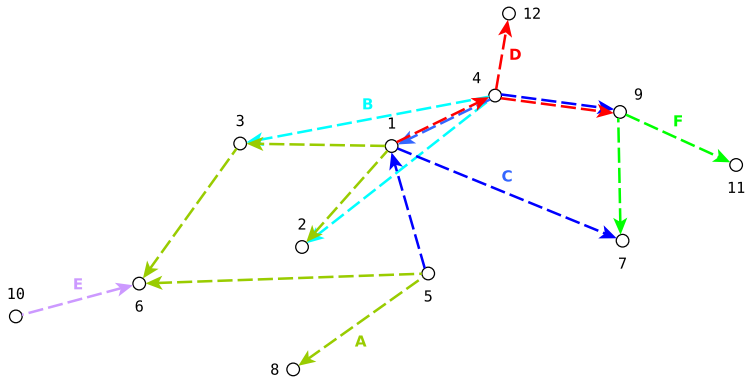


Figure 2.3 – *Spreading trace from Table 2.1, with events rearranged, sorting by file, in chronological order (above) and corresponding spreading cascades (below). Each peer is represented by a node in the graph and each event is represented by a dotted arrow, connecting provider to client. Each file is represented by a color and arrows are colored accordingly. Timestamps are not directly represented in this directed acyclic graph, though the chronology of the events can be found following the edges of the cascade.*

The first key property encoded in the spreading cascade of a given file F is the number of nodes who possess it at the end of the observed period, which is given by the *size* of the cascade $|\mathcal{P}_F|$. We also explore two other key topological

properties of the cascade, namely its *depth* and *number of links*. The former is defined as the length of the longest path on the cascade and captures the maximum number of hops from peer to peer that the file has undergone before it was relayed from a provider to a client. The number of links, given by $|\mathcal{L}_F|$, combined with the size of the cascade gives information on the sharing pattern of the network. For example, in Figure 2.3, the corresponding cascade to the file A has size 6, depth 2 and 5 links.

From the P2P trace log we have constructed the spreading cascades for each observed file and computed the above mentioned features. The distribution of these cascade features is presented in Figure 2.4. First, we observe that the cascade depth distribution is well fitted by a power-law. Examining individual cascades with high depth we realize that they are not typically big in terms of size. Second, most spreading cascades are quite small, featuring one or few nodes and links – these cascades are essentially trivial trees. The cascades with higher number of links, however, display a richer structure. In fact, the ones with the highest number of links cannot be tree-like, since their number of links exceeds (by far) the maximum cascade size observed in our dataset.

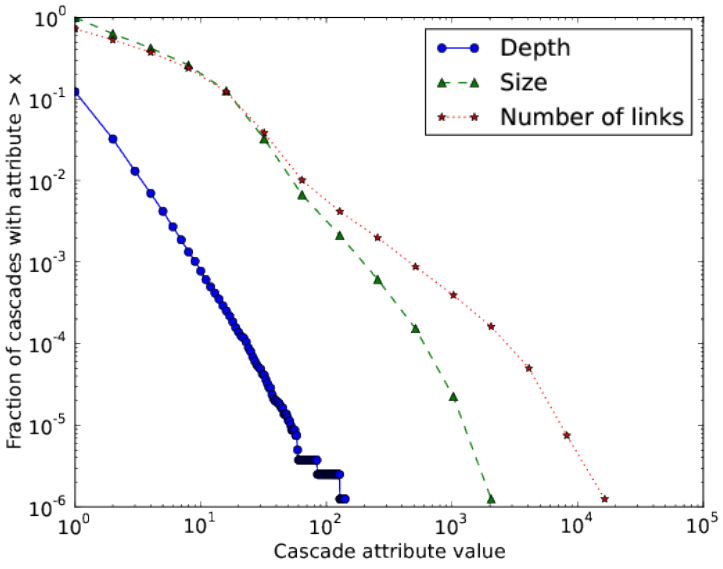


Figure 2.4 – *Complementary cumulative distribution of key properties (depth, size, links) of real spreading cascade.*

2.2.2 Initial providers

Another relevant spreading data concerns the *initial providers* or *seeds* for each file F, namely the set of peers that possessed it prior to any transfer activity on the observed trace. These nodes are the origin of the spreading cascades, triggering the diffusion of the file F. This information can also be inferred from the request log and be determined in the following way. Let $\mathcal{C}_F(t) = \{C \in \mathcal{P} : (t', \cdot, C, F) \in \mathbf{D}, t' < t\}$ be the set of peers who requested F prior to t. We define the set of initial providers of F as the set of peers P who have provided F at some time t, without having obtained it before t from another peer in the network:

$$\mathcal{I}_F = \{P \in \mathcal{P} : (t, P, \cdot, F) \in \mathbf{D}, P \notin \mathcal{C}_F(t)\}$$

To illustrate this concept, consider the spreading trace in Table 2.1: the set nodes of each spreading cascade corresponding to a file can be partitioned into a set of initial providers and another of clients:

File	Clients	Seeds
A	2, 3, 6, 8	1, 5
B	2, 3	4
C	1, 7, 9	4, 5
D	4, 9, 12	1
E	6	10
F	7, 11	9

Table 2.2 – *Spreading cascade nodes partitioned into seeds and clients: sample trace from Table 2.1.*

Plotting the complementary cumulative distribution of the number of initial providers for the spreading cascades (Figure 2.5) we obtain an interesting curve, revealing a scale-free distribution. This means that although most spreading cascades in our observation have few initial providers, there is a non negligible fraction of cascades with a large number of initial providers.

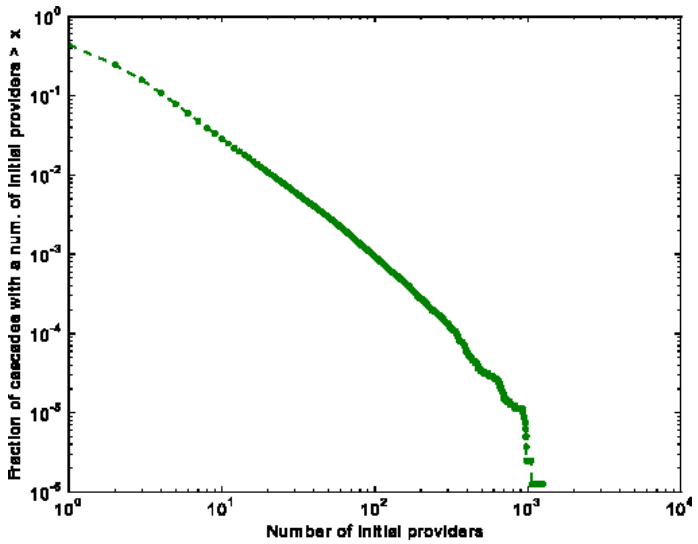


Figure 2.5 – *Complementary cumulative distribution of the number of initial providers.*

2.3 Underlying network

As discussed in the introduction, our goal is to investigate and model spreading cascades on the social network of peers participating in the P2P system in question. In order to analyze the empirical spread of files among peers in the light of detailed network diffusion models mentioned, we need not only the detailed chronological data of who transmitted the information to whom (observable in the trace) but also the social network on which the diffusion takes place. As pointed out in [Gomez-Rodriguez et al., 2012] it is challenging to reconstruct the network on which the diffusion takes place.

Focusing on content diffusion among peers, it is natural to consider the *interest graph* in which each node represents a peer and each edge joining two peers stand for common interest. Interests connecting peers may include broad subjects such as open source software, folk rock or French literature or narrower ones such as movies by Quentin Tarantino, a particular computer game or pictures of Beijing. It is reasonable to suppose that peers store and share content related to their interests and, likewise, peers will search for content matching their interests. Hence the diffusion of files among peers takes place on the interest graph and occurs from neighbor to neighbor. Indeed, if a peer P provides a file F (corresponding to a music album for example) to another peer P' then there is a link between

them in the interest graph, since both are interested in the same content, namely F.

One strategy to unfold this network in our context is to explore relations among peers and their common shared files. Such strategy was hinted in [Handurukande et al., 2006] and developed more substantially in [Latapy et al., 2008, Iamnitchi et al., 2011, Bernardes et al., 2012]. We follow this approach to reconstruct the underlying social network as well.

2.3.1 Bipartite structure of the data

The trace $\mathbf{D}$ captures directly a relationship between files and peers who share them. A natural way to organize these relationships is through a *bipartite graph* $\mathcal{B} = (\mathcal{P}, \mathcal{F}, \mathcal{A})$, a graph defined by two disjoint sets of nodes $\mathcal{P}$ and $\mathcal{F}$ and a set of links $\mathcal{A} \subset \mathcal{P} \times \mathcal{F}$ between a node in one set and a node in the other set. In our case, we construct the bipartite graph with the disjoint sets of all peers and all files in our data and for each recorded event in $(t, P, X, F) \in \mathbf{D}$ we add a link to $\mathcal{A}$, connecting the file F to the peers P and X, that is:

$$\mathcal{A} = \{(P, F) \in \mathcal{P} \times \mathcal{F} : (\cdot, P, \cdot, F) \in \mathbf{D} \vee (\cdot, \cdot, P, F) \in \mathbf{D}\}$$

where $(\cdot, \cdot, P, F) \in \mathbf{D}$ represents a recorded event in which some peer provided the file F to the peer P at some point in time and, likewise, $(\cdot, P, \cdot, F) \in \mathbf{D}$ represents a recorded event in which P provided the file F to some peer at some point in time.

In other words, $\mathcal{B}$ is the bipartite graph in which peers are linked to the files which they have provided or sought. The degree of peers and files in this bipartite graph represents the number of files transfered by a peer and the number of peers who shared a file, respectively.

As mentioned in the previous section, the degree of peers and files in this bipartite graph represents the number of files transfered by a peer and the number of peers who shared a file, respectively. Thus, we can relate it to the file sharing observations made in the beginning of the chapter. Indeed, as Figure 2.6 confirms,

the degree distribution of both peers and files is heterogeneous and mostly concentrated on small values with all degree values for peers and files remain below 10^4.

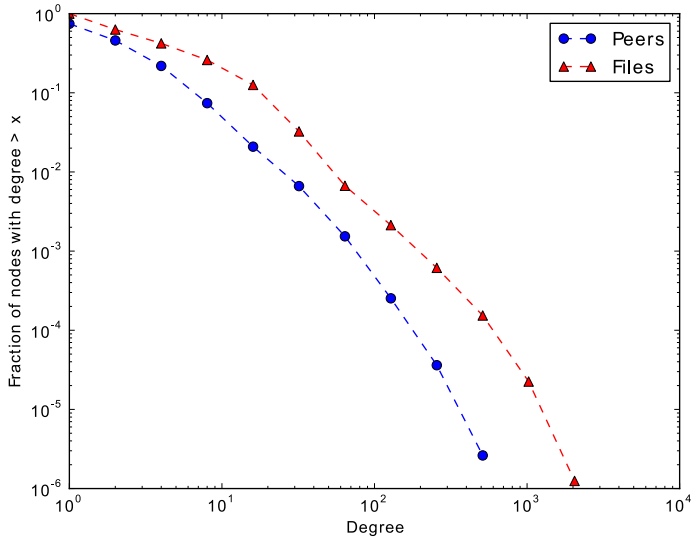


Figure 2.6 – *Complementary cumulative degree distributions of peers and files on the bipartite graph $\mathcal{B}$.*

2.3.2 Interest graph

It is beyond doubt extremely difficult in a large scale interaction network to know precisely whether any two individuals have a common interest. From the information encoded in $\mathcal{B}$ it is possible to draw relationships between the peers, projecting the bipartite graph on the set $\mathcal{P}$ [Diestel, 2010]. The projected graph $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ consists of a set of nodes $\mathcal{P}$ (the set of peers) and a set of links between these nodes $\mathcal{E}$, defined in the following way: two peers are connected if they have at least one neighbor in common (in $\mathcal{F}$) in the bipartite graph, that is:

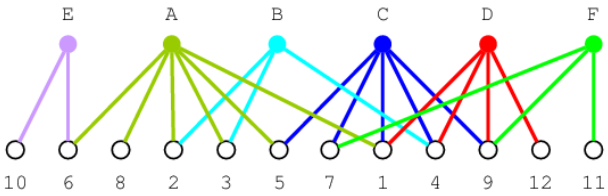
$$\mathcal{E} = \{(P, P') \in \mathcal{P} \times \mathcal{P} : \exists F \in \mathcal{F}, (P, F) \in \mathcal{A} \ \wedge \ (P', F) \in \mathcal{A}\}$$

This projection provides an approximation of interest graph described in the introduction of the section [2], for it connects any two peers who have manifested a common interest during our observations. We give in Figure 2.7 an illustration

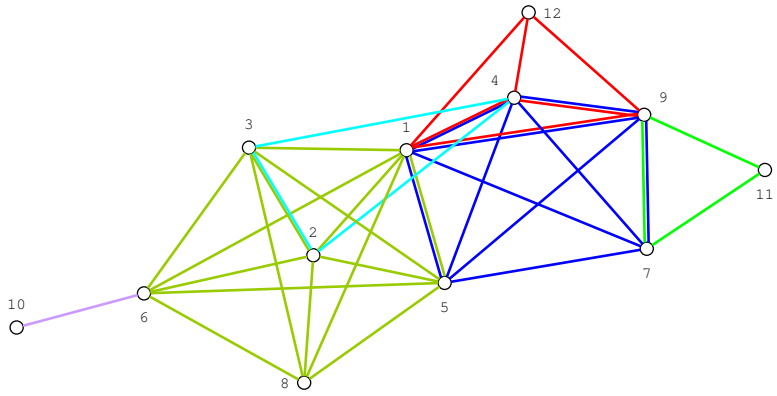
2. For the sake of readability the approximated interest graph will be henceforth denoted simply *interest graph*.

of this method applied to the sample trace given in Table 2.1. We first construct the bipartite graph of peers and files using the trace. Secondly, we obtain the interest graph projecting the bipartite graph in the set of peers. Notice that, by construction, the spreading of files takes place in the interest graph and occurs from neighbor to neighbor.

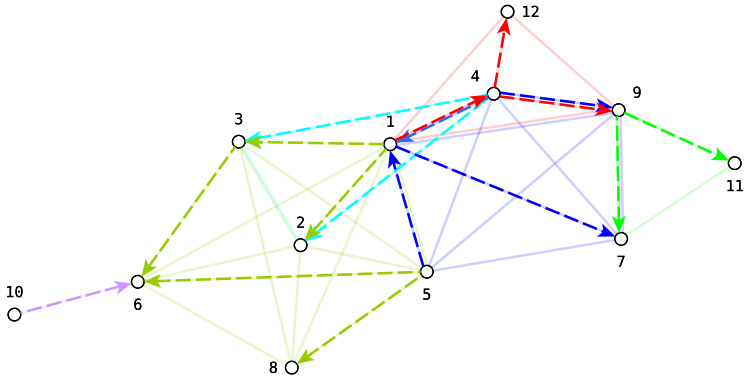
The interest graph obtained from the observed bipartite graph (as explained above and in Figure 2.7) has a single giant connected component containing essentially all nodes (99.99%), density 2.62×10^{-4} and diameter 13. In Figure 2.8a we have plotted the degree distribution for the peers: considering the set of all peers, the median degree is 118 and the mean value is 500.11, with corresponding standard deviation of 1271.42. We proceed to a finer analysis of the degree distribution, grouping peers in categories (Figure 2.8a). Let us consider first the set of *clients* $C \in \mathcal{P}$ such that $(\cdot, \cdot, C, \cdot) \in \mathbf{D}$: i.e., peers having requested files during our measurements. Their degree distribution superposes the degree distribution of all nodes. This is due to the fact that 99.63% of peers in our observations have requested at least one file, so the clients degree distribution is essentially the global degree distribution. A much more restrictive category is the set of *providers* P such that $(\cdot, P, \cdot, \cdot) \in \mathbf{D}$, i.e., peers having supplied files during our measurements. Their degree distribution has a similar shape, but it is concentrated on larger values, indicated by a median of 1821 and an average degree of 2906.54 – with corresponding standard deviation of 3471.80. The last curve, superposing the curve corresponding to the providers, represents the degree distribution of the initial providers. We have also computed the clustering coefficient (See chapter I for a discussion and definition) of the peers in the interest graph (Figure 2.8b): we observe a wide range of clustering values, each represented by a significant fraction of peers. Also, the distribution shows a relatively high fraction of peers with a high clustering coefficient – which is a feature of real complex networks, in contrast to random graphs.



(a) *Bipartite graph constructed from sample trace in Table 2.1, featuring 6 files (top) and 12 peers (bottom).*

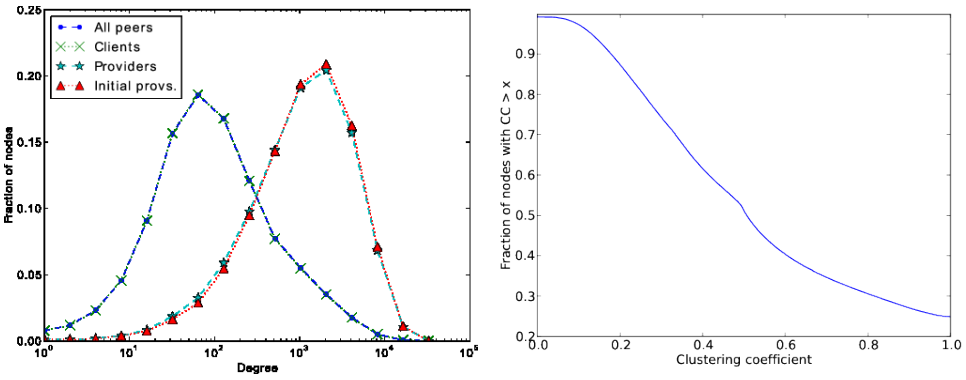


(b) *Interest graph as the projection of the bipartite graph above.*



(c) *Sample spreading cascades (Figure 2.3) superposed on interest graph.*

Figure 2.7 – *Interest graph construction and relation to the spreading cascade.*



(a) *Degree distributions on the interest graph. Superposed curves: all peers and clients, providers and initial providers* **(b)** *Complementary cumulative clustering coefficient distribution in the interest graph.*

Figure 2.8 – *Interest graph structural properties*

2.4 Summary

We close this chapter with a brief summary: we obtained an open dataset containing a large-scale diffusion trace from file sharing in P2P systems. First we reported file sharing properties of this dataset found in peer-to-peer literature, namely a heterogeneous file sharing behavior among peers and an overwhelming presence of free-riders [Handurukande et al., 2006]. Secondly, we examined the diffusion cascades obtained from the trace and observed that spreading cascades are mostly trivial with a small proportion cascades featuring complex topological structure, also in agreement with the literature [Leskovec et al., 2007, Liben-Nowell and Kleinberg, 2008, Goel et al., 2012]. In particular, key properties of spreading cascades are heavy-tailed, with cascade depth distribution featuring a scale free distribution.

Third, we have introduced a framework to infer the interest graph of peers, on which the spreading of files takes place. This graph connects essentially all peers, which can be grouped in two categories: providers and clients. Most peers in our observations are clients, but only a small fraction supply files and there is a sharp distinction between clients and providers in terms of their degree distribution. The structural properties of the interest graph – namely diameter, degree distribution and local clustering – are congruent with the literature on complex networks, as

discussed in the previous chapter.

In sum, the obtained dataset is a legitimate candidate to study large-scale diffusion and it allows us to assess the pertinence of diffusion models since it provides detailed information on the spreading process and the underlying social network of peers.

Chapter 3

SIR Model and extensions

Contents

As discussed in the first chapter, epidemic/contagion models are ubiquitous in the literature to describe empirical data (from epidemic to viral marketing to P2P file spreading) and to generate artificial diffusion. In particular the network version of the SIR model has been used since it is relatively simple and analytic tractable asymptotically. In the literature, some authors have been able to select parameters for SIR models such that they could generate simulated cascades similar the real cascades they have observed. In the following we take one step forward, by assessing if the spreading model is compatible with the data, if we calibrate the model assuming the observed diffusion is described by the model in question. That is, instead of of extensively searching the parameter space of the models for interesting values, we use a standard framework to estimate the model parameters from the data. We then simulate the calibrated model to generate artificial cascades

which we compare to real cascades.

In this chapter being examining the standard, simple SIR model as a baseline model and explore natural extensions of this model which capture heterogeneities, particularly in terms of peer behavior and file popularity. We also introduce an affinity measure among peers and examine an extensions of this model which take this measure into account.

3.1 Simple SIR model

We begin examining simple SIR model, generating simulated cascades and comparing them with real ones to assess how realistic this model performs on the interest graph, in terms of the following cascade properties: size, depth and number of links. Note that by realistic, we mean able to reproduce the characteristics of the data as we measured. [1]

In our setting the SIR model dynamic is as follows: each file spreading corresponds to an independent epidemic in the interest graph, in which each node is in one of the following states: *susceptible*, *infected* or *non-interacting* (sometimes named *removed*, hence the acronym SIR). Susceptible nodes do not possess the file and may receive it from an infected node, thus becoming infected. Each infected node, in turn, spreads the file to each of its neighbors, independently, with probability p and becomes promptly non-interacting thereafter. Although non-interacting nodes remain in this state, infected nodes may unsuccessfully try to infect them.

Supposing the observed diffusion trace is the result of such a simple SIR epidemic we may estimate the spreading parameter p. Each neighbor-to-neighbor

1. The problem of improving the measurement process is different from the one of identifying relevant models able to capture the features observed in the data, which is our focus. Indeed, our goal is to assess the ability of the models to reproduce (or not) the characteristics of real traces as *observed* in the dataset, reproducing the eventual shortcomings of the data. Sampling improvements include the application of detection techniques (such as [Secan et al 2011]) in order to remove abnormal events from the raw data before using the modeling techniques discusses in this chapter. Although they could potentially improve the data, they are not essential at this point, thus, we leave this approach for further work.

transmission trial can be seen as a Bernoulli random variable, whose value is 1 in case of success and 0 otherwise and whose expected value is p. Assuming each trial is independent and the parameter p is homogeneous for each P and F, we may estimate it by the empirical proportion of successes over all trials. Since each tuple in $\mathbf{D}$ accounts for a successful neighbor-to-neighbor transmission, $|\mathbf{D}|$ is the number of successful trials for all diffusion cascades. The total number of trials, in turn, is given by the sum of the degrees of all nodes involved in the spreading of each file. Hence, we obtain the following estimate, with a 95% confidence interval $\hat{p} \pm 10^{-6}$:

$$\hat{p} = |\mathbf{D}| \,/ \sum_{F \in \mathcal{F}} \sum_{P \in \mathcal{P}_F} d(P) = 1.063 \times 10^{-3}$$

Since the simple SIR model depends on a single parameter, namely the spreading probability p, we have fully characterized it with the preceding estimation.

3.1.1 Calibration

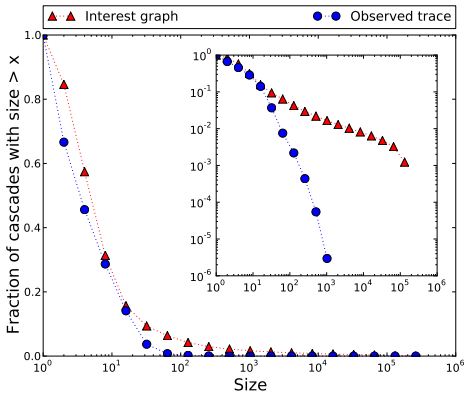
In this Section we use the reconstructed underlying network and the initial condition information (the list of initial providers $\mathcal{I}_F$ computed for each file F), obtained in the previous chapter, and the SIR model with calibrated spreading parameter $\hat{p}$, as described above, to simulate file spread diffusion. For each F, we begin with the initial providers in an infected state and the other nodes in a susceptible state. At each step, infected nodes infect each of their neighbors with probability $\hat{p}$, becoming non-interacting afterwards. The epidemic continues as long as there are interacting infected nodes.

The first observation concerning the model simulation is that the observed time (measured in seconds) has no direct relation with the simulation time (number of steps). Furthermore, our dataset corresponds to an observation in a bounded window of time of eight hours, so that we have no reason to suppose that the file spreading cascades we observe correspond to the whole spreading cascade of a file. In other words, if we had measured a longer time window we would likely observe bigger cascades (in terms of size and depth) for the same files –

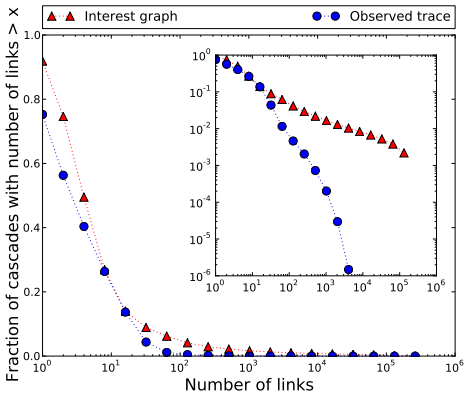
due to, among other reasons, new users who could eventually request the same files. This is also true for our SIR model: we observe increasingly bigger cascades as simulation time increases. In fact, performing unconstrained simulations we have obtained a distribution of significantly bigger cascades than the ones we have observed in the real trace. Thus, in order to perform a suitable comparison with the observed cascades, we have decided to hold one property fixed and compare the other properties. More precisely, for each file we generate a simulated cascade with the same size (resp. depth) as the corresponding observed cascade and compare the depth (resp. size) and number of links. In practice, for each file we simulate the SIR epidemic as described earlier and halt it when it reaches the size (resp. depth) of the corresponding observed cascade. We have performed $801\,280$ file spreading simulations in total (one for each file in $\mathcal{F}$).

3.1.2 Results

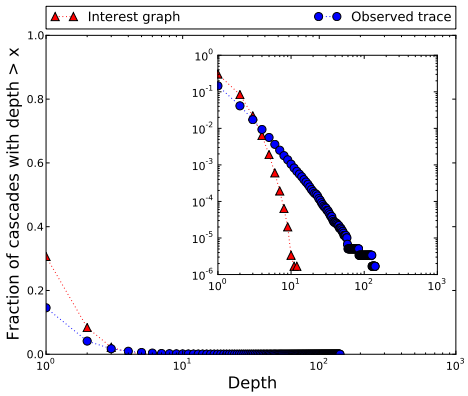
In Figure 3.1a we plotted the complementary cumulative distribution of the size of cascades with comparable depth. We observe a divergence of the cascade size from the observed cascades: simulated cascades are typically much bigger in size for a given depth compared to real cascades. The range of values in both categories is also striking: the biggest real cascade is at least two orders of magnitude smaller than the biggest simulated ones. In Figure 3.1c we plot the complementary cumulative distribution of the depth of cascades with fixed size. Real cascades feature a much higher depth compared to simulations, holding cascade size constant. In particular there is a cutoff on the cascade depth for the simulations: we do not observe any simulated cascade with depth bigger than 11. As for the number of links, we have two interesting situations. If we fix the depth (Figure 3.1b) the number of links distribution resembles closely the size distribution (Figure 3.1a). This is not completely surprising, since the two quantities are related. In this case we observe a larger number of links for all simulations compared to the number of links in the real cascades since the simulated cascades themselves are bigger. If, in contrast, we fix the cascade size to fit the observed cascades size (Figure 3.1d), we observe a typically smaller number of links. Combining these observations on both plots we conclude that real spreading cascades are denser than simulated ones, a clear qualitative feature not captured by the simple SIR model. Finally we note



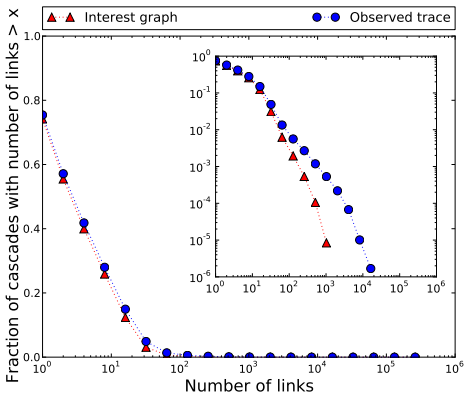
(a) *Size of cascades with fixed depth.*



(b) *Number of links of cascades with fixed depth.*



(c) *Depth of cascades with fixed size.*



(d) *Number of links of cascades with fixed size.*

Figure 3.1 – *Complementary cumulative distribution of key cascade properties for real and simple SIR-generated cascades.*

that most cascades are trivial, featuring depth equal to one and correspondingly small size.

To sum up, we have compared simple topological properties of real spreading cascades and simulated cascades from a calibrated SIR model, with comparable depth and size. We have observed that simulated cascades are relatively "wider" whereas real cascades are relatively "elongated", that is, real cascades have a smaller size per depth ratio. Moreover, real cascades are typically denser than simulated ones.

3.2 Heterogeneous SIR models

In the previous section we have examined the adequacy of the simple SIR model to generate realistic file spreading cascades. Given the generality and simplicity of the homogeneous model, it is not entirely surprising that it does not capture key properties of real spreading cascades in our data. In order to fairly assess the relevance of the SIR dynamic in our context, in this Section we consider natural extensions of the SIR model considered previously, which take into account heterogeneous aspects found in the observed data. More precisely, we perform a complementary analysis, focusing on the interest graph and examining two heterogeneous versions of the SIR model, characterized by a distribution of spreading probabilities, instead of a single homogeneous parameter. These models take into account the file popularity and peer behavior heterogeneity and are, thus, presumably better equipped to mimic real spreading cascades.

3.2.1 File popularity

A first refinement of the simple SIR model consists in introducing different spreading probabilities according to the file being spread. The rationale in this case is to account for different levels of popularity depending on the file. Exogenous reasons – such as a movie release or the death of an artist – can change the supply and demand of a given file and consequently alter its spreading probability. If we know the spreading probabilities for each file, i.e., $\{p(F) : F \in \mathcal{F}\}$, the knowledge of the actual reasons that explain the heterogeneity in file popularity are irrelevant to the characterization of this model. An estimate of these probabilities, in turn, can be obtained from the trace $\mathbf{D}$ if we suppose it was generated by a process following this extended SIR model. Indeed, since each file spreading is independent of the others, it is possible to estimate $p(F)$ for each F separately, with the same method used to derive the homogeneous parameter. Restricting the calculations to the spreading cascade of F, $\hat{p}(F)$ will be given by the empirical proportion of successful transmissions of F over all possible transmissions of F:

$$\hat{p}(F) = |\{(\cdot, \cdot, \cdot, F) \in \mathbf{D}\}| \,/\, \sum_{P \in \mathcal{P}_F} d(P)$$

In Figure 3.2a we plot the distribution of the heterogeneous spreading parameters depending on the files. The values of $\hat{p}$ are concentrated on the range 10^{-5} to 10^{-2}, indicating that there is a considerable fraction of cascades with a significantly different spreading regime (bigger than one order of magnitude). This distribution characterizes the extended SIR model we use in the following simulations.

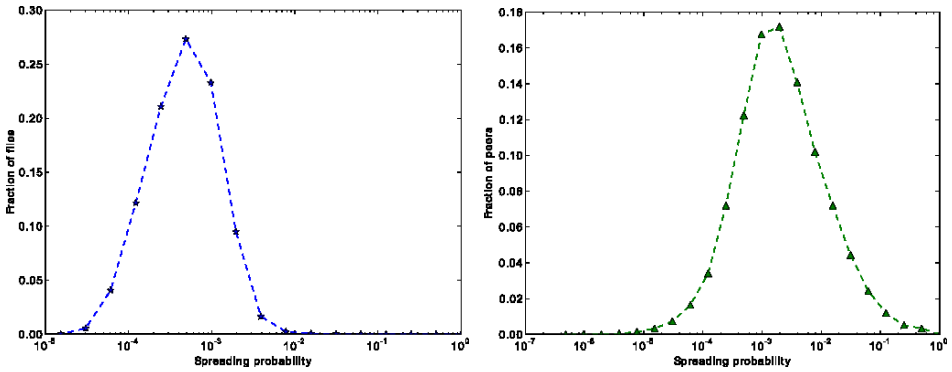
3.2.2 Peer behavior

A second possible refinement is motivated by the fact that peers might have intrinsically distinct levels of "generosity" regarding file sharing. Under this hypothesis we extend the standard SIR model assigning an heterogeneous spreading probability to each peer, regardless of which file it is sharing. Thus, we do not need any other information but the spreading probability distribution to characterize the model. In this context altruistic peers, who typically spread files to a large proportion of their neighbors, would feature a bigger spreading probability compared to the homogeneous spreading probability corresponding to the diffusion aggregates of all peers. By the same token, the extreme case of free-riders would have their spreading probability assigned to zero. Again we can study transmissions as outcomes of Bernoulli trials to estimate the spreading probabilities. Let $\mathcal{F}_P = \{F \in \mathcal{F} : (P, F) \in \mathcal{A}\}$ be the files carried by the peer P; for each such file the number of transmission trials P could perform corresponds to its degree in the interest graph, namely $d(P)$. Hence, to obtain $\hat{p}(P)$ for each peer P we divide the number of successful transmissions of P to other peers (of any file carried by P) over the total number of potential trials:

$$\hat{p}(P) = \frac{|\{(\cdot, P, \cdot, \cdot) \in \mathbf{D}\}|}{|\mathcal{F}_P| \times d(P)}$$

We have plotted the distribution of the positive spreading probabilities estimates in this case (Figure 3.2b). They account for small fraction of all the peers, since the only peers who have a positive spreading probability are those who provided a file at least once – namely 4.33% (cf. observations made in Chapter 2). Conversely, a large fraction of the peers do not share the file in this model. We observe a marked range of values, which is significantly greater than the one

computed for the homogeneous SIR.



(a) *Depending on the files* **(b)** *Depending on the peers*

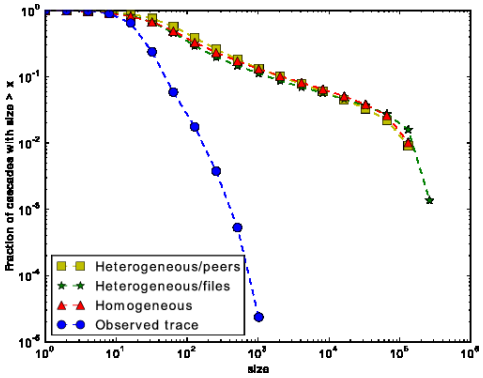
Figure 3.2 – *Heterogeneous spreading parameter distributions*

3.2.3 Results

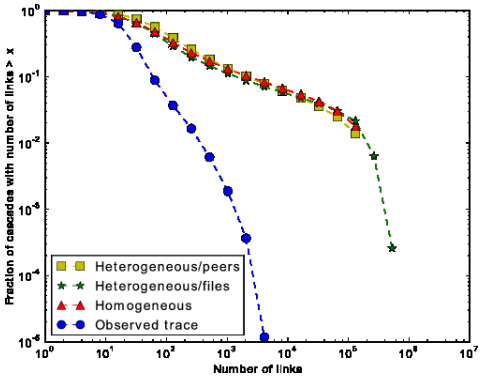
Our aim is to generate simulated cascades following both extensions of the SIR model presented – with heterogeneous spreading probability depending on the files and on the peers – and compare their properties with simulated cascades of the simple SIR model and the real observed cascades. In this sense, we apply the same methodology as in previous simulations: we fix the depth (resp. size) for the simulated cascades and examine the other two properties – the idea is to compare similar spreading cascades in terms of the chosen property. As discussed previously, the great majority of the cascades is simple, with depth equal to one and a small size. Hence the simulated cascades corresponding to the simple observed cascades will likely correspond in terms of depth, size and number of links. For this reason, we have decided in this Section to focus on the spreading cascades with depth greater than one.

The simulation results are plotted in Figure 3.3: we have plotted the complementary cumulative distributions of the spreading cascade depth, size and number of links. Imposing a constraint on the depth for the simulated cascades and comparing their size (Figure 3.3a) we observe the contrast between the simulated and the real observed cascades with the same depth: the former have a typically bigger

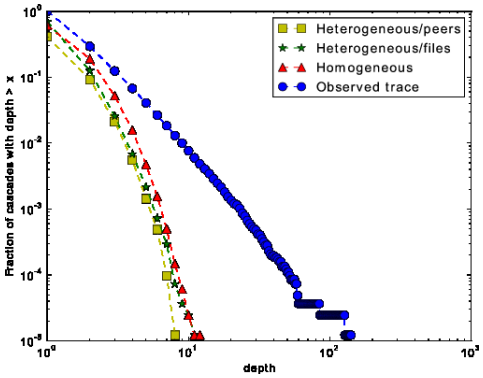
size compared to the latter. What is remarkable, however, is the agreement among all the simulated cascade distributions – curves superposed in Figure 3.3a. Next, if we fix the size for the simulated cascades and examine their depth (Figure 3.3c), we face the same qualitative similarity among simulated curves. Indeed, the curves corresponding to the heterogeneous SIR models also feature a cutoff in depth, failing to reproduce the scale-free curve representing the depth of the observed real cascades. Finally, the cascade links distribution plotted in Figure 3.3b and Figure 3.3d confirms the pattern observed previously, namely that the observed spreading cascades are typically denser than corresponding simulated cascades.



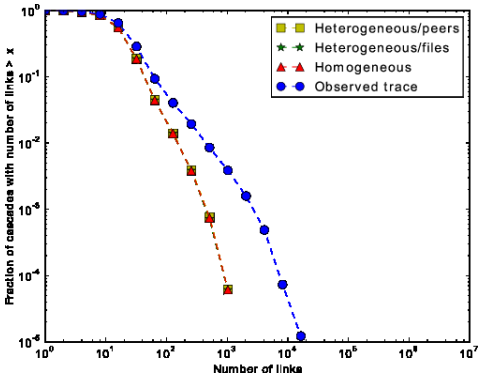
(a) *Size of cascades with fixed depth. Curves corresponding to the simulations are superposed.*



(b) *Number of links of cascades with fixed depth. Curves corresponding to the simulations are superposed.*



(c) *Depth of cascades with fixed size.*



(d) *Number of links of cascades with fixed size. Curves corresponding to the simulations are superposed.*

Figure 3.3 – *Simulation of file spreading on the interest graph with heterogeneous SIR extensions: complementary cumulative distribution of cascade properties.*

In spite of the improvements in the SIR model, introducing an heterogeneous spreading parameter to account for different profile of files (respectively peers), simulations indicate that this refinement does not change qualitatively the basic properties of simulated spreading cascades. Indeed we observe a surprising similarity between the three compared SIR models, notwithstanding the particularities of each model.

3.3 SIR model with affinity measure

In the previous Section we have examined SIR model extensions that take into account heterogeneous aspects of peers and files with the goal of generating more realistic spreading cascades. Another approach is to keep the simple SIR model and enrich the social network inference. In this Section we address this question, proposing a way to refine the interest graph taking into account the *affinity* among peers. The rationale is that peers are more likely to interact with other peers with whom they have greater affinity. In the following we describe a method to quantify this relation.

3.3.1 Weighted interest graph

In concrete terms, our affinity score between two peers will be defined by the number of common files peers shared or provided. Indeed, instead of approximating the interest graph by the simple projection of $\mathcal{B}$ on $\mathcal{P}$, we consider a richer inferred interest graph $\mathcal{G} = (\mathcal{P}, \mathcal{E}, \mathcal{W})$, given by the *weighted* projection of $\mathcal{B}$ on $\mathcal{P}$ such that

$$\mathcal{E} = \{(P, P') \in \mathcal{P} \times \mathcal{P} : \exists F \in \mathcal{F}, (P, F) \in \mathcal{A} \wedge (P', F) \in \mathcal{A}\}$$

$$\mathcal{W}(P, P') = |\{F \in \mathcal{F} : (P, F) \in \mathcal{A} \wedge (P', F) \in \mathcal{A}\}|$$

In other words, peers belonging to the neighborhood of a common file in $\mathcal{B}$ are connected in $\mathcal{G}$. If a peer P provides a file F (corresponding to a music album for example) to another peer P', then there is a link between them in the interest graph since both are interested in the same content, namely F. Furthermore, each

edge $(P, P') \in \mathcal{E}$ has an integer weight given by the number of common files they have manifested interest in. As an example, consider the trace sample from Chapter 2: the corresponding weighted interest graph, reproduced in Figure 3.4, will take into account the "multiple" colored edges connecting two nodes in the Figure 2.8a, reproduced below. Thus, the edges $(1, 4), (1, 5), (1, 9), (2, 3), (4, 9), (7, 9)$ have weight 2 and the other edges have weight 1.

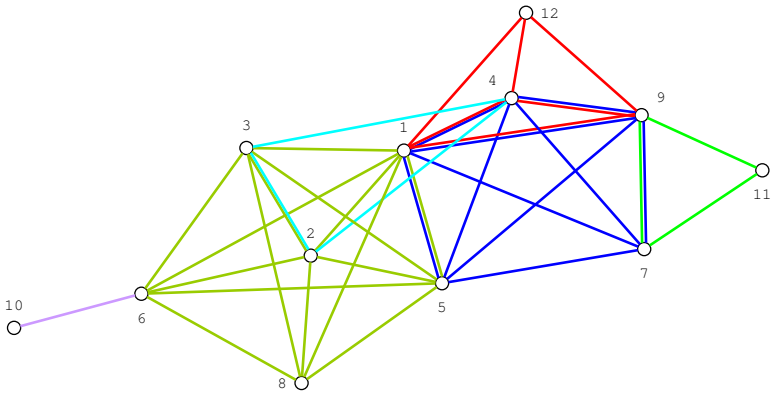
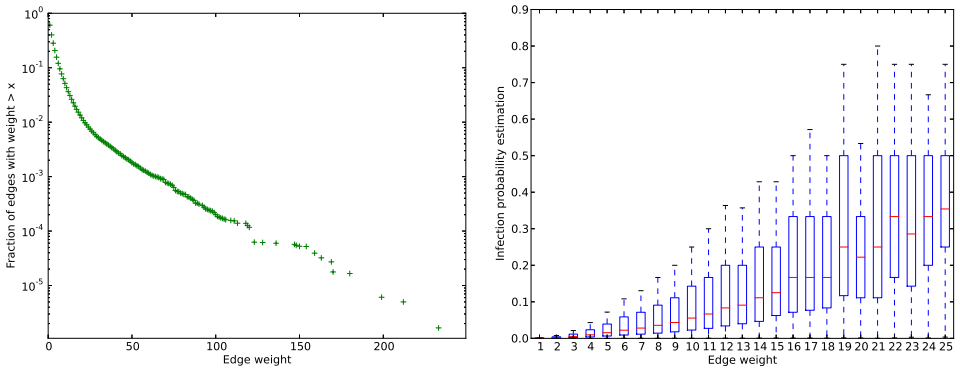


Figure 3.4 – *Weighted interest graph reconstructed from the simple trace given in Table 2.1.*

In Figure 3.5a we have plotted the distribution of weight values in the interest graph: it is heterogeneous, with the vast majority of edges featuring small weights. Finally, note that the weight scheme we have introduced is by no means the only way to assign an affinity index to each edge of the interest graph. One could assign a greater affinity to two peers who are both interested in rarer files than two peers interested to common files for instance; another possibility is the Jaccard index of similarity. That said, our choice is quite natural and is motivated by the hypothesis that peers will likely spread files to the neighbors with whom they have greater affinity, as we explain below.

3.3.2 Affinity measure and spreading dynamic

The diffusion models we have used so far require adaptation to take into account the enhanced network topology. We keep the main hypotheses of the SIR model, that is, that each individual is in one of the following states: *susceptible*,



(a) *Edge weight distribution in the interest graph: heterogeneous (heavy-tailed)*

(b) *Infection probability estimation, revealing an increasing spreading probability with weight*

Figure 3.5 – *The interest graph connects peers who share common interests and attributes a weight between this connection proportionally to the the overlap among their interests. Some peers have several common interests with others, but most peers have few shared interests. Contagion spreads best among peers with stronger connection.*

infected or *non-interacting* (sometimes denoted *removed*). Susceptible nodes do not possess the file and may receive it from an infected node, thus becoming infected. Infected nodes, in turn, try to spread the file to each of its neighbors, independently, and become promptly non-interacting thereafter. Each infection attempt from an infected node P to the node P' is successful with probability $\sigma(w) \in [0, 1]$, depending on the weight w of the edge connecting P and P'.

It is reasonable to assume that a peer P is more successful in spreading a file to the neighbors with whom he or she has a greater common interest. In terms of the spreading probability σ, this assumption translates itself as supposing $\sigma(w)$ is increasing with w. Indeed, the weight connecting P and its neighbors is a measure of how similar are their interests. Hence the more similar two peers are in terms of interest, the greater the weight of the edge connecting them and, in turn, the greater the spreading probability. To verify this hypothesis we have estimated the value of $\sigma(w)$ for each value of w, adapting estimation methods used in Sections 3.1, 3.2. Each observed spreading cascade of a file F in the trace provides a set of estimated values $\{\hat{\sigma}_F(w)\}$: as expected, we have found that the median values of $\hat{\sigma}$ are increasing with w up to $w = 25$ (with the exception of two values), after which they essentially reach a plateau at $\hat{\sigma}(w) = 0.5$. In Figure 3.5 (right) we

have plotted the estimator values for all weights from 1 to 25 in terms of box plots.

Following the approach in [Onnela et al., 2007], we have used a linear function to model the spreading probability on the weighted graph, namely $\sigma_1(w) = a_1 w + b_1$, with $a_1 = 3.07 \times 10^{-3}$ and $b_1 = 1.54 \times 10^{-3}$ obtained with a least squares calibration. The number of edges with small weights is much greater than the number of edges with big weights in this graph – cf. Figure 3.5a. Indeed we observe a greater number of transmissions between peers connected by edges with smaller weight. Hence, the quality of the estimators is greater for small values of w and we have taken into account primarily these values in this model. We have also examined an alternative model for σ, which captures qualitatively the stagnation of σ for large values of w. In this case we have $\sigma_2(w) = a_2 \log(w) + b_2$ with $a_2 = 14.10 \times 10^{-3}$ and $b_2 = 0.58 \times 10^{-3}$ obtained with the same calibration method.

3.3.3 Results

Equipped with the reconstructed social network of peers (the weighted interest graph) and models for the diffusion of files (described above) we have simulated the spreading of all the files and compared the corresponding spreading cascades with the real, observed, spreading cascades. Simulated traces corresponding to the spreading of each file $F \in \mathcal{F}$ contains the same number of transfers as the real observed trace of F.

In Figure 3.6 we have plotted the complementary cumulative distributions of cascade properties from real cascades, compared to the simulated cascades using the diffusion models described above. The first general remark is that simulated cascades generated by both models are quite similar in terms of these metrics. Indeed, the curves of both simulations are superposed for the three plots. Compared to the distribution of real cascades, the sharpest contrast is in terms of depth: the distribution for simulated cascades features only small values of depth, whereas the depth distribution for real cascades is remarkably scale-free. We also find a discrepancy between simulated and real cascades in terms of size and number of links: in the former the gap is sharper and in the latter both distributions follow

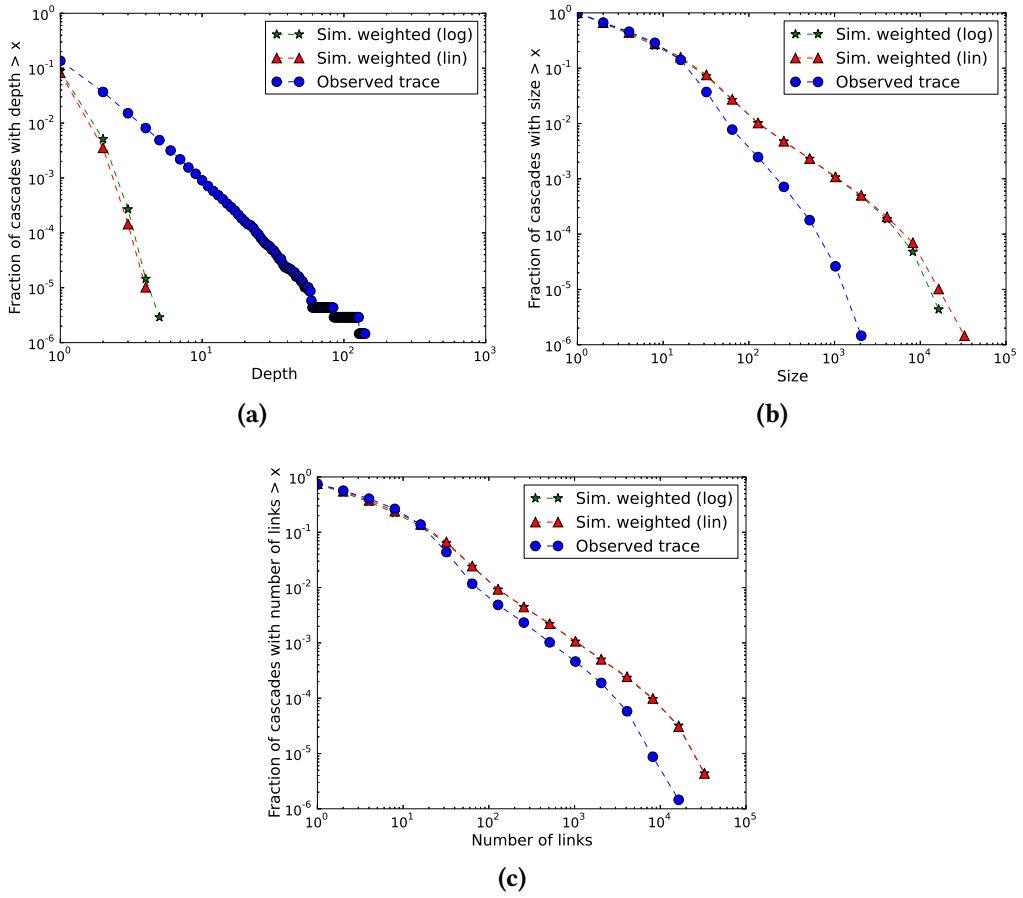


Figure 3.6 – *Spreading cascades profile in terms of depth, size and number of links respectively. Both models yielded the same cascades profile (simulation curves superposed), contrasting with real spreading cascades in terms of depth.*

globally the same trend. Considered together the curves make clear that these models face a challenge to capture key topological properties simultaneously. Indeed, real cascades have a shape closer to chain-email cascades [Liben-Nowell and Kleinberg, 2008], in the sense that they are relatively elongated compared to simulated cascades obtained with these contagion models.

3.4 Summary

We have assessed the pertinence of SIR model and extensions, using a maximum of likelihood estimation framework. Assuming the observed diffusion was

a product of an epidemic contagion process, we have calibrated the models and generated simulated cascades which we compared to real ones. We concluded that simulated file diffusions do not capture key qualitative properties of the observed spreading cascades.

Simulated cascades from extensions of the SIR model (which take into account the heterogeneity in file popularity and peer behavior) show similar properties as the simple homogeneous SIR model. In addition to these extensions, we have enriched the reconstruction of the interest graph, introducing a measure of affinity among peers. Again, simulations reveal another unexpected point: despite the enhanced social network topology, the model simulations did not reproduce qualitative features of real spreading cascades.

The reason behind these results may be that the SIR model is too simple to account for the diffusion mechanism. Although this is a likely possibility, it is remarkable that taking into account the above mentioned heterogeneities did not improve the model significantly. This suggests that the key component to improve the model is other. As we shall see in the next chapter, integrating time patters into this process is a hopeful strategy to improve models.

Chapter 4

Temporal patterns

Contents

In the previous chapter we have developed a model calibration and evaluation framework and began assessing the simple SIR model. Once we concluded it was incapable of reproducing key spreading cascade properties, we examined several extensions of the model which explored several properties found in the data, both in terms of the diffusion process (taking into account file popularity and peer behavior) and in terms of the underlying network (weighted interest graph). In particular, in the latter extension peers who were not much active in the network, exchanging only a few files, had a small affinity score with their neighbors (since their affinity score is limited by the number of files). Thus, files spread more difficultly to these peers, relative to more active/present peers in the system. As the introduction of the affinity measure did not led to any significant qualitative improvement, we concluded it failed to capture heterogeneity of node presence in the network properly.

With the goal to take into account the node presence directly, we consider the *dynamic interest graph* of peers, which is obtained from the structure of the original

interest graph in addition to the intervals of presence of each node. As mentioned in Chapter 2 we do not dispose of the peer connection data in our dataset. Thus we begin this chapter describing a method to infer the connection instants for each peer using their activity pattern (which contains temporal data, in terms of timestamps). Next, in order to study diffusion on the dynamic interest graph, we need to departure from the simple SIR dynamic and examine models capable of taking into account this temporal information in the spreading mechanism. More precisely, since the interval connections are given in terms of seconds, we need a diffusion model whose evolution is given in a compatible time scale. We motivate the choice of such a model and corresponding adaptations it entails in our framework. Similarly to the previous chapter, we perform simulations to assess the capability of this model (in two variations) to reproduce realistic cascade properties.

4.1 Peer connection data

Although we do not dispose of connection events for each peer in our dataset, we know the activity pattern of each peer (in time), as our dataset consists of a collection of file exchange records among peer with timestamps. We summarize peers' behavior in the file sharing system in Figure 4.1: since we only record transmission events, each observed peer has connected at least once into the system and remained a certain time on-line during our measurements.

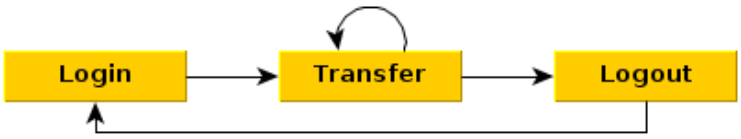


Figure 4.1 – *Scheme of peer activity featuring all possible events.*

Intuitively, given the activity profile, such as the examples in Figure 4.2, we would like to place connection and disconnection events in the timeline. In order to do so in a systematic way we first have to make a few assumptions on peer behavior concerning connection events. A simple model is to suppose that connection and disconnection times occur after exponential times and that the time

elapsed between two file request events is relatively short when users are on-line and longer if users went off-line. In this case, given the activity pattern of a node, we can infer if the time elapsed between two file requests are of long type or short type using an expectation–maximization algorithm, as well as the exponential rates [Jewell, 1982]. If we dispose of this information and we know that peers likely wait a "typical" amount of time from the moment they connect into the system and the first file request we can obtain a likely connection instant. Analyzing a similar dataset of P2P request collected in our lab we have determined that the typical time in this context was 5 minutes. In Figure 4.2 we illustrate the method, showing the inferred connection events for the peers featured in Figure 2.1. With this procedure we estimated the distributions for the login and logout rates: Figure 4.3 shows the complementary cumulative distributions for the estimated peer login and logout rates and observe they are heavy tailed.

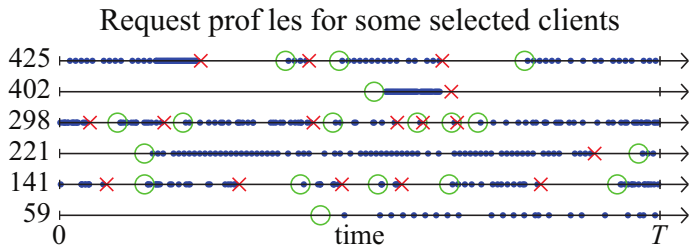


Figure 4.2 – *Activity profile of six clients with inferred connection events: green circles and red crosses represent logins and logout respectively.*

4.2 Integrating time patterns

Once the connection data for the peers has been obtained, in the following we use the data to improve the interest graph of peers, defining the *dynamic interest graph*, where peers interaction is only taken into account if peers are simultaneously present in the system. This new interest graph also calls for new spreading models, which are able to take into account the connection data, thus we examine an adapted version of the network SI model, in two variations: one which considers a homogeneous peer spreading behavior and another which features an individual spreading behavior for each peer.

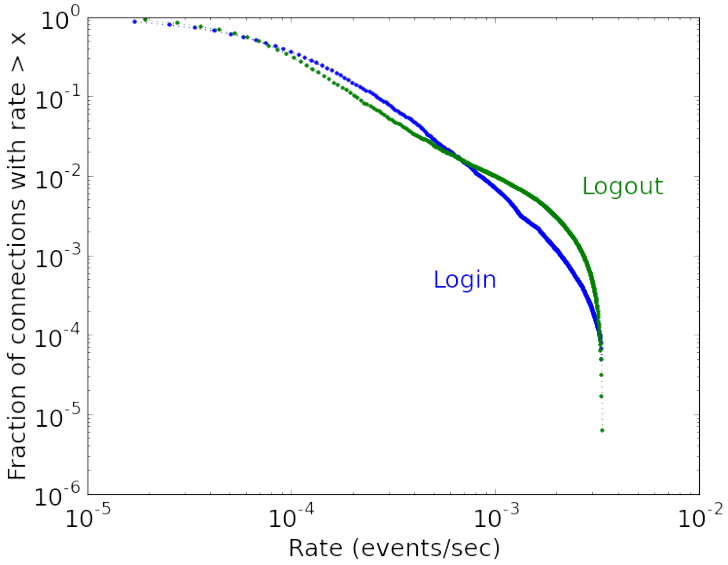


Figure 4.3 – *The peer login and logout rate complementary cumulative distributions are also heterogeneous and feature a cutoff.*

4.2.1 Dynamic interest graph

The interest graph is a comprehensive synthesis of peers' interest relations revealed in the observed time window. These relations are key to diffusion, since the spread of files occurs on the interest graph, as pointed out previously. However, even if the spread of files between neighbors in the interest graph is likely, the actual transfer of files may not occur concretely because they may never be simultaneously connected to the P2P system or have a small *co-presence time* – i.e., the amount of time on-line in the presence of each other in the system is small. Hence, in order to make simulations more realistic, in the sense of reproducing observed file spreading cascades, we used temporal information to enhance the social network reconstruction.

A strategy to use temporal information, integrating the connection data estimated in the previous section is to reconstruct a *dynamic interest graph*. In this graph, two peers will be connected at time $t > 0$ if they share a common interest (as in the interest graph) and if they are both online at time t. More formally, let $\mathcal{P}_t$ be the set of nodes on-line at time $t > 0$ and let the dynamic interest graph be

defined as $\mathcal{G}_t = (\mathcal{P}_t, \mathcal{E}_t)$, with

$$\mathcal{E}_t = \{(P, P') \in \mathcal{P}_t \times \mathcal{P}_t : \exists F \in \mathcal{F}, (P, F) \in \mathcal{A} \text{ and } (P', F) \in \mathcal{A}\}.$$

Intuitively, the dynamic interest graph is built similarly to the original interest graph, but evolves with the addition/suppression of connecting/disconnecting nodes and the respective links between these nodes and their neighbors. The dynamic interest graph is a subgraph of the interest graph $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ defined previously, in the sense that for all $t > 0$, $\mathcal{P}_t \subset \mathcal{P}$ and $\mathcal{E}_t \subset \mathcal{E}$. In the following, we examine the dynamic interest graph as the underlying social network on which we perform file spreading simulations.

4.2.2 Spreading dynamic with inter-contagion time

In the previous chapter we have modeled the spread of files using a SIR model in which nodes are in one of the following states: *susceptible*, *infected* and *removed*. A node in the latter state is permanently inactive and cannot infect other neighbors. In this chapter, where we explore and take into account peers' temporal patterns, the inactive periods correspond to the off-line periods, encoded in the dynamic graph. Thus, to simulate the file spreading, we use the SI model, a contagion model similar to the SIR model. In this model, each individual is either *susceptible* or *infected*. Susceptible nodes do not possess the file and may receive it from an infected node, thus becoming infected. Infected nodes, in turn, try to spread the file to each of their neighbors in the network, one at a time.

In this model we also introduce a new feature: the time between two infections takes a random number of seconds following an exponential distribution, which we refer to as the *inter-contagion time (ICT)*. Node latency given by exponential time is a common assumption and was proposed previously in the context of P2P file sharing systems [Leibnitz et al., 2006]. The ICT is characterized by a rate or, alternatively, by the mean (expected) ICT, since in the case of exponential random variables the mean time is the inverse of the rate. Moreover, if P possesses the file F, the number of peers who received the file F from P (after P obtained it) is a

Poisson process characterized by the inter-contagion time rate (or the mean ICT). We will examine SI models with *homogeneous* and *heterogeneous* inter-contagion time. In other words, in the first case we suppose all nodes have the same spreading behavior (global ICT rate) and in the second, an individual one (a different ICT rate for each node).

The introduction of contagion model featuring inter-contagion times allows us to adjust the simulation in terms of the chronological time (in seconds), as we observe in the diffusion trace. This represents a key contrast to the spreading models from the previous chapter, whose evolution happened in a simulation intrinsic time (given by the number of steps in the algorithm). This is precisely the reason why we had to hold one cascade property constant and analyze the remaining properties in the previous chapter: in this way we would have a comparable set of cascades with respect to a property. In contrast, the time bound of the simulations using the model presented above is given in seconds, so there is a more straight-forward and natural way to obtain a comparable set of simulated cascades: we impose the same time scale observed in the diffusion trace to the simulated cascades. That is, we simply simulate the diffusion of the cascades up to the time T (last time observed in the trace) and compare the three key properties of the simulated cascades to the corresponding real ones.

With the methodology presented above, we proceed to the model calibration, using the temporal data in our trace. The estimation process takes into account the number of files provided by each node and how long the node was on-line. Therefore, it yields different estimates for the average inter-contagion time in the static and dynamic settings – i.e., if we suppose nodes were continuously on-line during the whole period or not. Considering the homogeneous SI model first, we estimate average inter-contagion times of $10\,064$ seconds (2h48min) in the static setting and $4\,926$ seconds (1h22min) in the dynamic setting.

Next, considering the heterogeneous SI model, we also have different average inter-contagion time estimates for different settings: similarly to the homogeneous model, individual estimates are also generally greater in the static setting. Indeed, nodes seem less active if we suppose they were continuously on-line in the whole

observation period (since the number of transfers remains the same). An important difference in this model, compared to the homogeneous one, is the following: individual average inter-contagion times imply that observed free riders (clients who do not provide files) have null ICT rate estimates. Hence they will also behave as free riders in simulations of this model. The estimated complementary cumulative distributions in both settings (static and dynamic) are plotted in Figure 4.4. As noted in Chapter 2, more than 95% of the peers in the system are free riders, and thus, are not represented in the plot.

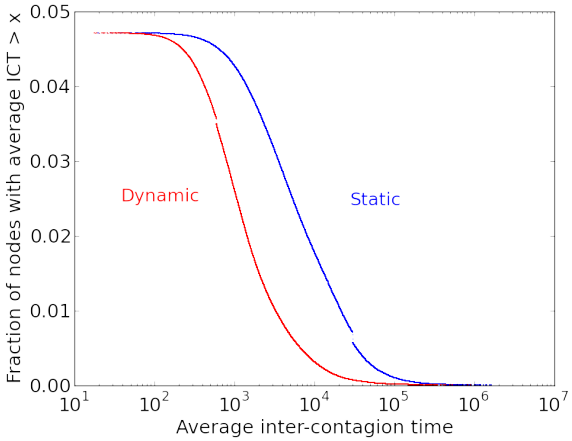


Figure 4.4 – *Complementary cumulative distributions of individual average inter-contagion time estimates for nodes in the static and dynamic interest graphs. Free riders ($> 95\%$) have null inter-contagion time rate and are not shown.*

4.3 File spreading simulation

We have simulated the SI model with homogeneous and heterogeneous spreading behavior as outlined above on the dynamic interest graphs for each file present in the trace. The profiles of real and simulated cascades are summarized in Figure 4.5: we have plotted the complementary cumulative distributions of cascades' size, number of links and depth. For each cascade property, we plot the same distribution in lin-log and log-log (inset) scales, which highlight respectively smaller/short cascades (most cascades) and bigger/deeper cascades (rare cascades).

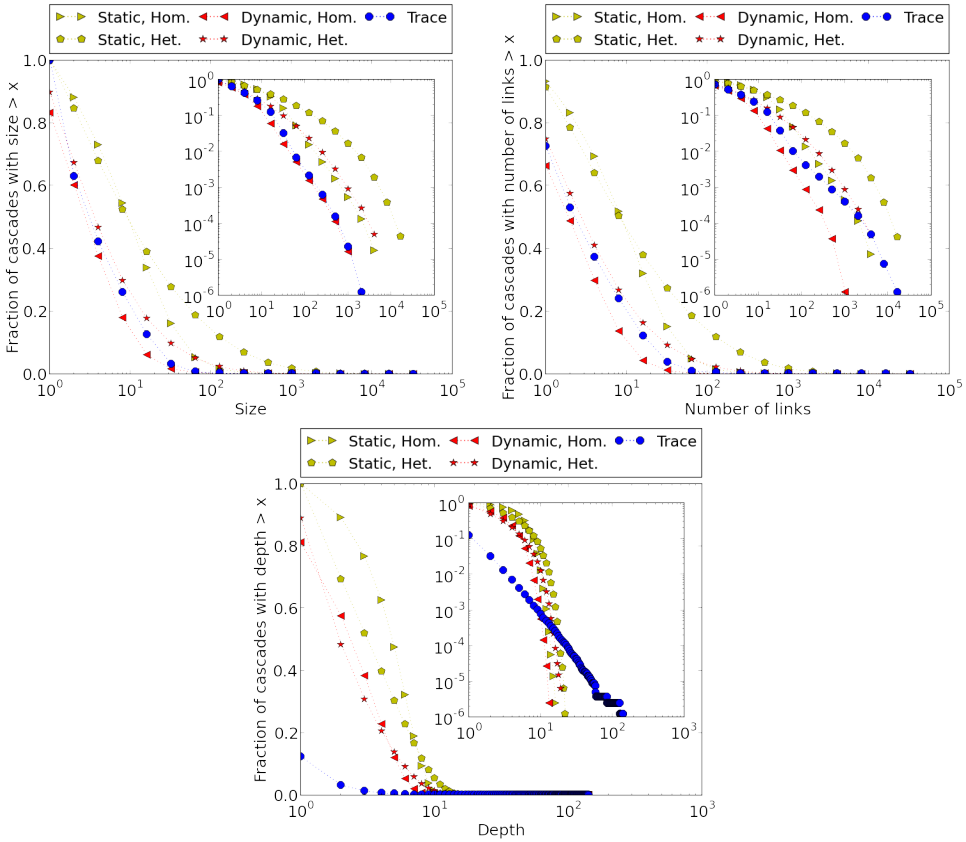


Figure 4.5 – *Spreading cascades profile in terms of size, number of links and depth, respectively. Plots feature the complementary cumulative distribution of these properties in lin-log and log-log (inset) scales. Simulations on the dynamic graph remain closer to real cascades (trace), with the homogeneous model reproducing well real cascades' size and the heterogeneous one, their number of links; no model was able to reproduce the observed depth distribution.*

4.3.1 Result

In terms of the variations examined, the dichotomy static/dynamic graph changes has an overall impact, but affects particularly the distribution of trivial cascades. Compared to simulations on the static graph, simulations on the dynamic graph yielded a bigger proportion of small cascades which is what we observe in out measurements. The dichotomy homogeneous/heterogeneous SI model impacts mostly the properties' distribution tail, particularly in terms of size and number of links. All other things equal, in our setting, the homogeneous SI model yielded simulations with smaller proportion of large cascades, which is closer to the observed cascades in terms of size, but more distant in terms of number of nodes. In

terms of depth distribution we note that none of the proposed models was able to reproduce the scale-free depth distribution featured by the real cascades: simulated cascades exhibit, in contrast to real ones, a sharp decrease in the proportion of cascades with depth greater than 10, revealing a cutoff.

In sum, in terms of size and number of links, we find encouraging results: both homogeneous and heterogeneous models perform relatively well in the dynamic setting, in the sense that simulations on the dynamic graph feature a proportion of small cascades similar to the real ones (most cascades). In terms of larger (and infrequent) cascades, the homogeneous model reproduces well the size distribution of real cascades; in terms of number of links, the heterogeneous model is superior. Although this model cannot generate artificial cascades similar to real ones in terms of all key properties, we have shown the importance of taking into account the temporal data in contagion models which aim to generate realistic cascades.

4.3.2 Impact of on-line presence

With respect to the previous approach, in Chapter 3 we have changed both the underlying network and the spreading dynamic, since it had to be compatible with the dynamic graph in terms of time scale. The converse, however, is not true: the new spreading dynamic we have used can be simulated in static graphs, particularly the original interest graph. Hence, we decided to simulate it also in this graph to isolate the impact of the new spreading dynamic from the the impact of the improved underlying graph. We denote in this chapter the original interest graph *static*, in contrast to the *dynamic* one. The results are quite different as we observe in Figure 4.5: comparing simulated cascade profiles on the static and dynamic interest graphs we note that cascades are generally smaller and feature a smaller number of links in the dynamic graph. As expected there are no properties for which the simulations on the least realistic graph were superior, though we insist that no model was able to reproduce the cascade depth distribution.

This can be due to the fact that this graph is static or that it simply ignores the fact that many connections could not the used in the real-world since peers were never simultaneously on-line. In our case, these “artificial” links which do

not respect the co-presence in the graph amount to 29% of the links in the static interest graph. So, in order to evaluate their impact, we have also simulated our models on the static interest graph without these links (not shown) and found that the impact was minor: simulated cascades in this new static graph featured the same profile of simulated cascades on the original static interest graph. Thus, we conclude that the difference in cascade profiles simulated on static and dynamic graphs is primarily due to the reduction of the *co-presence time* (and not simply the co-presence) among neighbors in the dynamic graph and potential causality effects. Indeed, in the static interest graph the co-presence times correspond to the whole observation period. This cascade profile difference is not trivial given the possibility that the co-presence time reduction could have been compensated (or overcompensated) by the fact that nodes in the dynamic graph are more active than nodes in the static graph, as discussed previously.

4.4 Summary

In this chapter we have explored the peer temporal patterns and their implications for epidemic contagion models such as those described in the previous chapter. First, given the activity profile, we have inferred the connection events of each user with a maximum of likelihood approach. Even though we were working with fairly simple starting assumption, namely that the connection intervals and the intervals between two requests followed an exponential distribution, with a rate per peer, we found that the rate distribution is heavy-tailed. This is due, in part, to the cutoff in the measurement in the end of our time window.

Secondly, we have adapted the interest graph to incorporate the connection data, thus constructing a dynamic interest graph. In addition to refining the underlying network, we have also improved the diffusion process, integrating the notion of inter-event times. In contrast to the models from previous chapter, this model add a latency in the information spread for each node. We have estimated the parameters for this model using the static and the dynamic interest graphs, supposing that peers have an homogeneous and heterogeneous behavior regarding this latency time. Simulation with these variants revealed that the most important

variation was the improvement in the interest graph. The difference between the static and dynamic interest graphs was key to reproduce realistic small cascades. Bigger cascades remain challenging to reproduce, specially in terms of depth.

In sum, the results of this chapter emphasize the value of integrating time patterns into the models, in order to generate realistic spreading cascades. In particular, we have highlighted the positive impact of considering dynamic graphs which integrate node connection data.

Chapter 5

The underlying network structure influence

Contents

Having analyzed the ability of contagion spreading models to reproduce key features of real file spreading cascades in the previous chapters, we turn to the question of the sensibility of these models to the underlying network structure. As we have mentioned in Chapter 1, there are a number of results in relating topological structures of random graphs and asymptotic results, i.e., when the epidemic is allowed to evolve with no duration constraints. In particular, studies have highlighted the importance of node degree distribution in random graphs to predict the probability of the epidemic extinction in the case of SIR models [Pastor-Satorras and Vespignani, 2001, Newman, 2003]. Similar results were also given in terms of graph spectral analysis [Wang et al., 2003, Prakash et al., 2012]. The analysis of asymptotic behavior of contagion models on random graphs featuring structural properties similar to real-world graphs is is an active research camp, particularly sparse random graphs with local clustering [Coupechoux and Lelarge, 2012]. In spite of the great interest and important results obtained in this area, the analysis of "out of the equilibrium" spreading, that is, in non-asymptotic regimes remains remarkably scarce – which is in part due to the challenges in devising

theoretical results without asymptotic analysis tools.

In this chapter we analyze the impact of interest graphs' key structural properties in terms of spreading cascades generated by the previous models. In particular we consider our baseline model, the simple SIR model and the best model examined, the SI model with homogeneous/heterogeneous inter-contag~ion time.

5.1 Methodology

As we have seen in Chapter 2, the interest graph was constructed from the measurements of real peer-activity and features key properties of complex networks, namely small diameter, heavy-tailed degree distribution and sparsity/local clustering. A priori, each of these properties (and perhaps other non-identified properties) may play a key role in contagion spreading. Moreover, it is possible that one property may have a much greater impact than another, independently of the other properties mentioned. For this reason these properties have to be analyzed separately. Strictly speaking this is not possible, since these properties are correlated. However, in the following we present a methodology to analyze these properties individually using random graphs.

We have considered the spreading of files in a sequence of random networks derived from the interest graph, with increasing topological complexity (Figure 5.1). More precisely we begin considering an Erdös-Rényi (ER) random graph with the same density and size as our interest graph, the simplest random graph in our sequence. Then we have chosen a random graph with the same density and degree distribution using the Configuration Model (CM) approach [Molloy and Reed, 1995, Newman, 2003]. Next we have generated a random bipartite graph, with the same density and degree distribution as our original bipartite graph $\mathcal{B}$ of peers and files [Guillaume and Latapy, 2004]. Compared to the interest graph, the projection of this random bipartite graph (RB) has similar density, degree distribution and clustering coefficient. In sum, for each new element of this sequence of (uniformly chosen) random graphs we introduce a new constraint to make it more realistic –

in the sense that its topological properties will be closer to the interest graph.

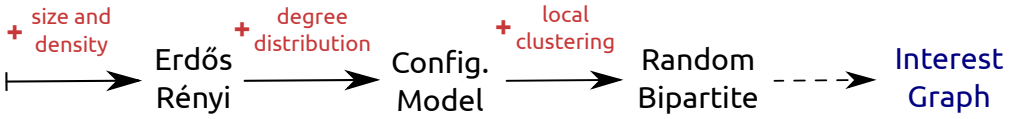


Figure 5.1 – *Increasingly realistic random graphs derived from the data, which replicate properties found in the interest interest graph. The random graphs were generated uniformly.*

In section 5.3, in which we simulate the best model investigated in the previous chapters, we examine an extra property in this section, in addition to the network structure variations provided by the sequence of random graphs: the degree distribution of the initial providers (sometimes also referred as epidemic "seeds"). As we have seen in Chapter 2, the observed degree distribution in the trace is different from the overall degree distribution of all nodes. Indeed, providers are typically more connected than regular nodes cf. Figure 5.2.

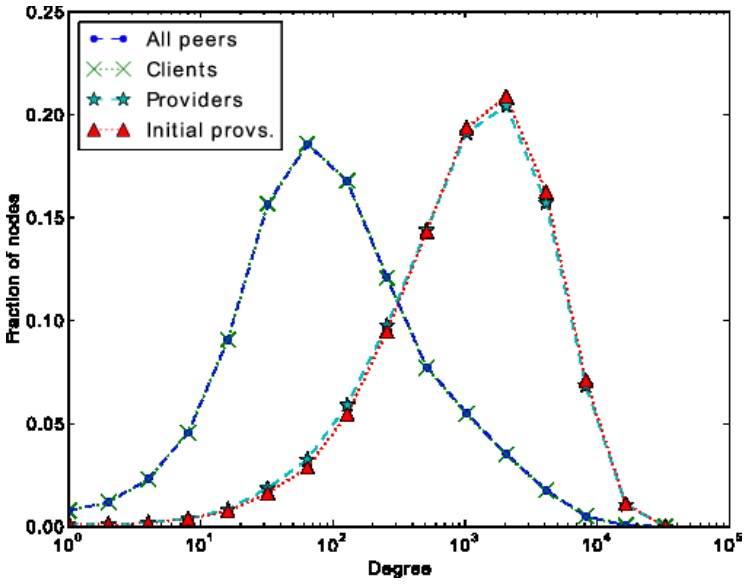


Figure 5.2 – *Degree distributions on the interest graph plotted in lin-log scale. Superposed curves: all peers and clients, providers and initial providers*

Hence, to assess the impact of the initial provider degree distribution we perform additional simulations on each random graph, holding this property constant. In the case of the ER graph, this extra simulation is not pertinent, since in these

graphs not even the node degree distribution is conserved. In the case of the CM graph, it is possible to simulate the model using precisely the same distribution observed in the trace. Finally, in the case of the RB graph, although the node degree distribution is similar to the interest graph's, one cannot guarantee a perfect match for each node. For this reason we ranked the nodes in the interest graph and in the RB graph in terms of their degree distribution and matched these nodes. The result is a similar and consistent degree distribution, suitable for to the simulation of all cascades.

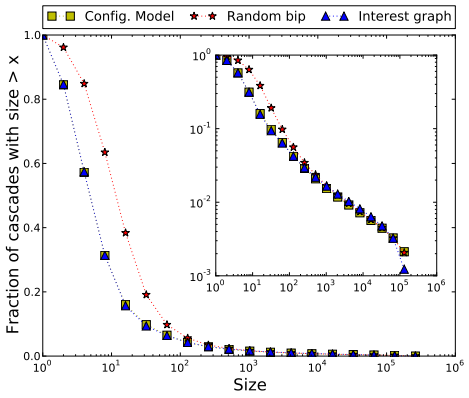
5.2 Simple SIR model on interest graph

As we pointed out in the beginning of this chapter, we are interested in investigating the impact of the key topological properties for contagion models in non-asymptotic regimes. Recall that in Chapter 3, we saw that the simple SIR model evolves in discrete simulation steps which have no direct relation with the the real time (measured in seconds) of real traces. Out of simplicity, we wish to compare the sets of spreading cascades similar to the real, observed cascades, so we will follow the strategy presented in Chapter 3: we identify the properties of each the observed cascades and generate simulated cascades with similar properties. In particular, we have decided to hold one property fixed and compare the other properties. More precisely, for each file we generate a simulated cascade with the same size (resp. depth) as the corresponding observed cascade and compare the depth (resp. size) and number of links. In practice, for each file we simulate the SIR epidemic as described earlier and halt it when it reaches the size (resp. depth) of the corresponding observed cascade.

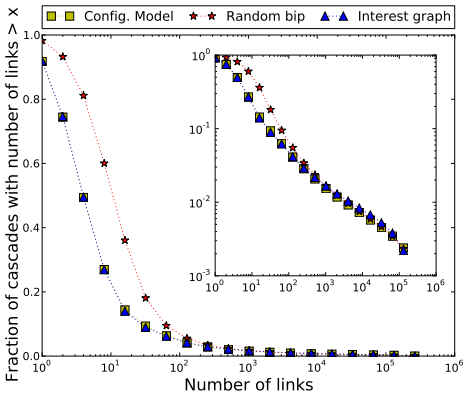
We have generated populations of simulated cascades for each underlying network and constraint (on depth and size). We have performed $801\,280$ file spreading simulations (one for each file in $\mathcal{F}$) for each network and have selected every simulated file spreading cascade which attained the depth (resp. size) of the real spreading cascade for the same file – and have rejected the others for purpose of comparison. With this procedure, each underlying network yields a different population of file spreading cascades, since the rejected cascades may be different

in each case. However 93.80% of the files have generated simulated cascades with the same depth as the corresponding real cascades, for all networks. Similarly, 85.64% of the files have generated simulated cascades with the same size as the corresponding real cascades, for all networks – except the ER network. Indeed, only 21.76% of the files have generated the contemplated size in the ER graph. Furthermore the properties of these simulated cascades on the ER graph deviated significantly from the properties of the cascades on the other graphs. Hence, in the following analysis we do not include the simulations for the ER graph. Rather, we focus on the properties of the files with comparable spreading cascade depth (resp. size) on all networks but ER.

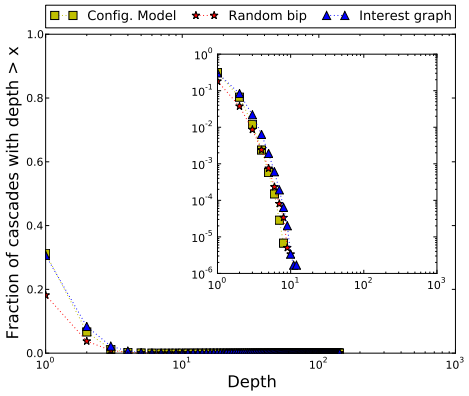
In Figure 5.3a we plotted the complementary cumulative distribution of the size of cascades with comparable depth. We observe a divergence of the cascade size from the observed cascades: simulated cascades are typically much bigger in size for a given depth compared to real cascades. The range of values in both categories is also striking: the biggest real cascade is at least two orders of magnitude smaller than the biggest simulated ones. Among the simulated cascades, there is a remarkable matching in size values for the simulation on the CM and the interest graph (curves are superposed). In Figure 5.3c we plot the complementary cumulative distribution of the depth of cascades with fixed size. Real cascades feature a much higher depth compared to simulations, holding cascade size constant. In particular there is a cutoff on the cascade depth for the simulations: we do not observe any cascade depth bigger than 11 in the simulations. As for the number of links, we have two interesting situations. If we fix the depth (Figure 5.3b) the number of links distribution resembles closely the size distribution (Figure 5.3a). This is not completely surprising, since the two quantities are correlated. In this case we observe a larger number of links for all simulations compared to the number of links in the real cascades since the simulated cascades themselves are bigger. If, in contrast, we fix the cascade size to fit the observed cascades size (Figure 5.3d), we observe a typically smaller number of links. Combining these observations on both plots we conclude that real spreading cascades are denser than simulated ones, a clear qualitative feature not captured by the simple SIR model. Finally we note that most cascades are trivial, featuring depth equal to one and correspondingly small size.



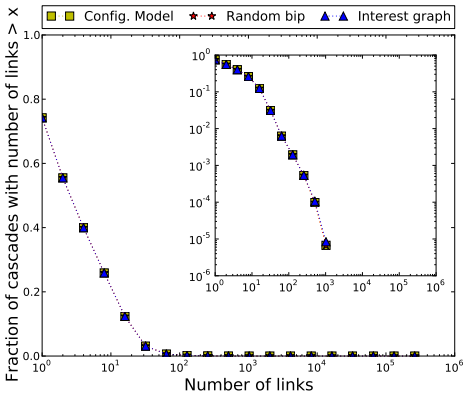
(a) *Size of cascades with fixed depth. Curves corresponding to the interest graph and CM superposed.*



(b) *Number of links of cascades with fixed depth. Curves corresponding to the interest graph and CM superposed.*



(c) *Depth of cascades with fixed size.*



(d) *Number of links of cascades with fixed size. Curves corresponding to the interest graph, RB and CM superposed.*

Figure 5.3 – *Simulation of file spreading on different underlying networks: complementary cumulative distribution of cascade properties*

To sum up, we have compared simple topological properties of real spreading cascades and simulated cascades from a calibrated SIR model, with comparable depth and size. We have observed that simulated cascades are relatively "wider" whereas real cascades are relatively "elongated", that is, real cascades have a smaller size per depth ratio. Moreover, real cascades are typically denser than simulated ones. In terms of interplay between underlying network structure and the simple SIR spreading cascades, we have observed that respecting the interest graph degree distribution was the only property that caused a striking change in simulations

behavior on the considered random networks. Indeed we have observed sharp qualitative dissimilarities between the simulations on the ER graph (different degree distribution) and no sensible dissimilarities between the simulations on the CM, RB and the interest graphs.

5.3 SI model on the dynamic interest graph

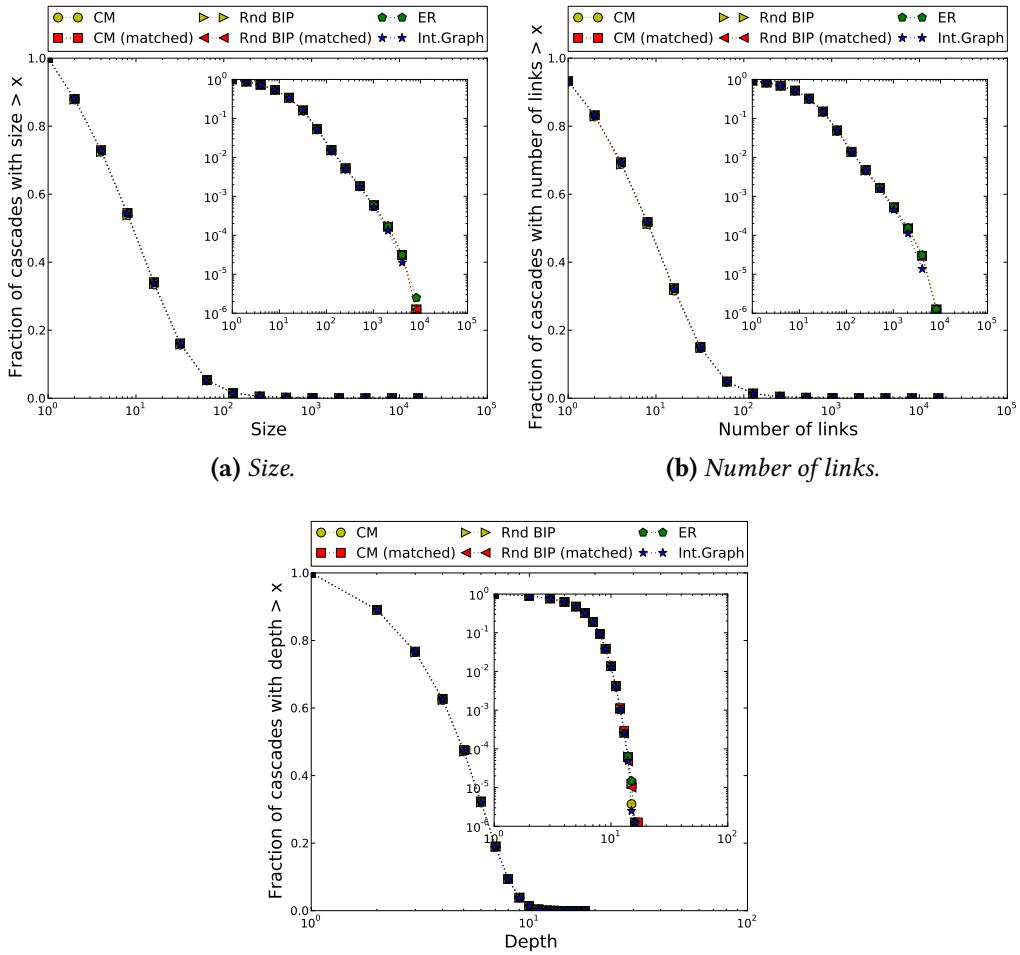
In our analysis, we proceed as in section 5.1, where we defined a sequence of increasingly realistic random graphs, in the sense that they have a topology increasingly similar to the interest graph. Recall the schematic representation of this graphs in Figure 5.1: we begin with an Erdös-Rényi (ER) random graph with the same density and size as our interest graph, the simplest random graph in our sequence. This graph is followed by a Configuration Model (CM) random graph with the same density and degree distribution. Next we have generated a random bipartite graph with degree distribution as our original bipartite graph, whose projection in the set of peers (RB) yields a graph with similar density, degree distribution and local clustering as the interest graph. In sum, for each new element of this sequence of (uniformly chosen) random graphs we introduce a new structural constraint to make it closer to the interest graph.

In the following, we simulate the spread of the files $\mathcal{F}$ on the random graphs described in the previous paragraph using the models examined in the Chapter 4, namely the SI with homogeneous and heterogeneous node behaviors (in terms of inter-contagion time distributions). For each SI model we perform two simulations: the first with the (static) random graphs and a second simulation in which we consider dynamic versions of the random graphs in the first simulation. More precisely, we have used the methodology to generate a dynamic interest graph from the static interest graph using the connection data, presented in section 4.2.1. At each instant $t > 0$ each node is present in the random graph if it was on-line at this instant in the P2P system (or equivalently, if it is present at this instant in the dynamic interest graph).

5.3.1 Homogeneous node behavior

We begin simulating the spread of the files $\mathcal{F}$ on the random graphs using the simplest model explored in this chapter, namely the SI with homogeneous inter-contagion time. The first batch of simulated cascades was generated using the following static graphs: ER, CM and RB graphs, with shuffled initial providers and CM, RB and the interest graph with matching initial providers. The results are plotted in Figure 5.4, where we see the six curves superposed for each property distributions plot. This indicates that the model is insensitive to all the variations we examined, which suggests that it is not a realistic model to capture user interaction. Surprising as this result may seem, it is not so different from the results obtained in the previous chapter, when we have investigated the impact of the topology for key cascade properties. Indeed, in that experiment, simulations on random graphs yielded similar sets of cascades for all graphs, except the ER graph, which failed to produce a comparable set of cascades in that framework. Compared to the models examined in the previous chapter, these models are different in the following aspects: it features an inter-contagion time and users remain active until they disconnect at the final time T. These are important differences, but when considered in isolation, they were insufficient to generate different sets of cascades. The same is true for heterogeneity: this change had no significant impact on the structure of the generated cascades.

In the following we examine the SI model and variations in the same random graphs tested previously, taking into account the connection patterns of the nodes. That is, each random graph is rendered dynamic, considering the connection/disconnection times for each node, as computed in the beginning of this chapter. We have generated a set of cascades using the homogeneous model and plotted the results in figure Figure 5.5. The generated sets of cascades remain similar to the sets of cascades generated in the previous trial, albeit with slight more variance between curves.



(a) *Size.*

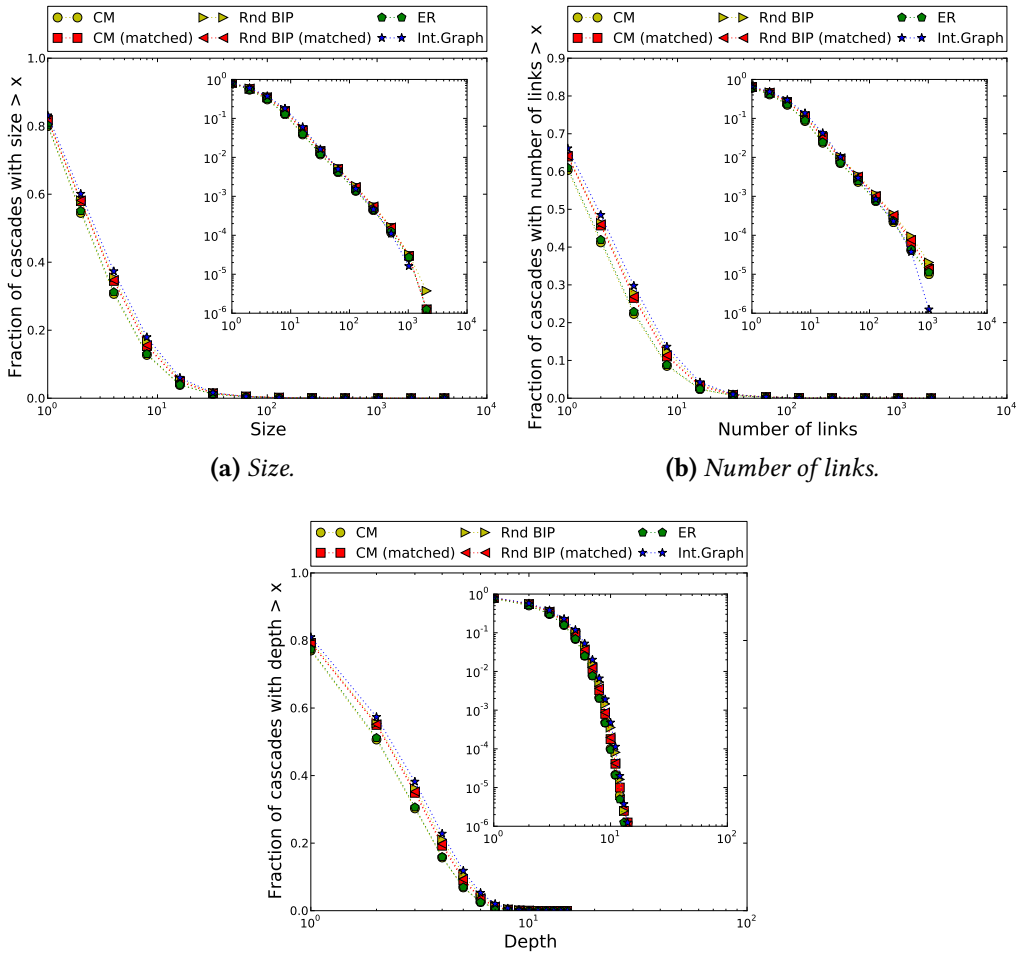
(b) *Number of links.*

(c) *Depth.*

Figure 5.4 – *Complementary cumulative distribution of cascade properties in static random graphs and homogeneous node behavior. Plots shown in lin-log and log-log scale (inset). Spreading model is insensible to variations on network topology and initial providers. All curves superposed in the three graphs.*

5.3.2 Heterogeneous node behavior

Now, we consider the same setting as the first trial, but using a SI model with a heterogeneous behavior of nodes. More precisely, we consider first the static random graphs (Figure 5.6), followed by the dynamic random graphs (Figure 5.7), as we did previously. In contrast to the previous simulations, we see a sharp distinction between two types of cascade profiles, indicated by the superposition of two sets of curves. That is, in the interest graph and in the random graphs where we have matched the degree distribution of the initial providers, the simulations



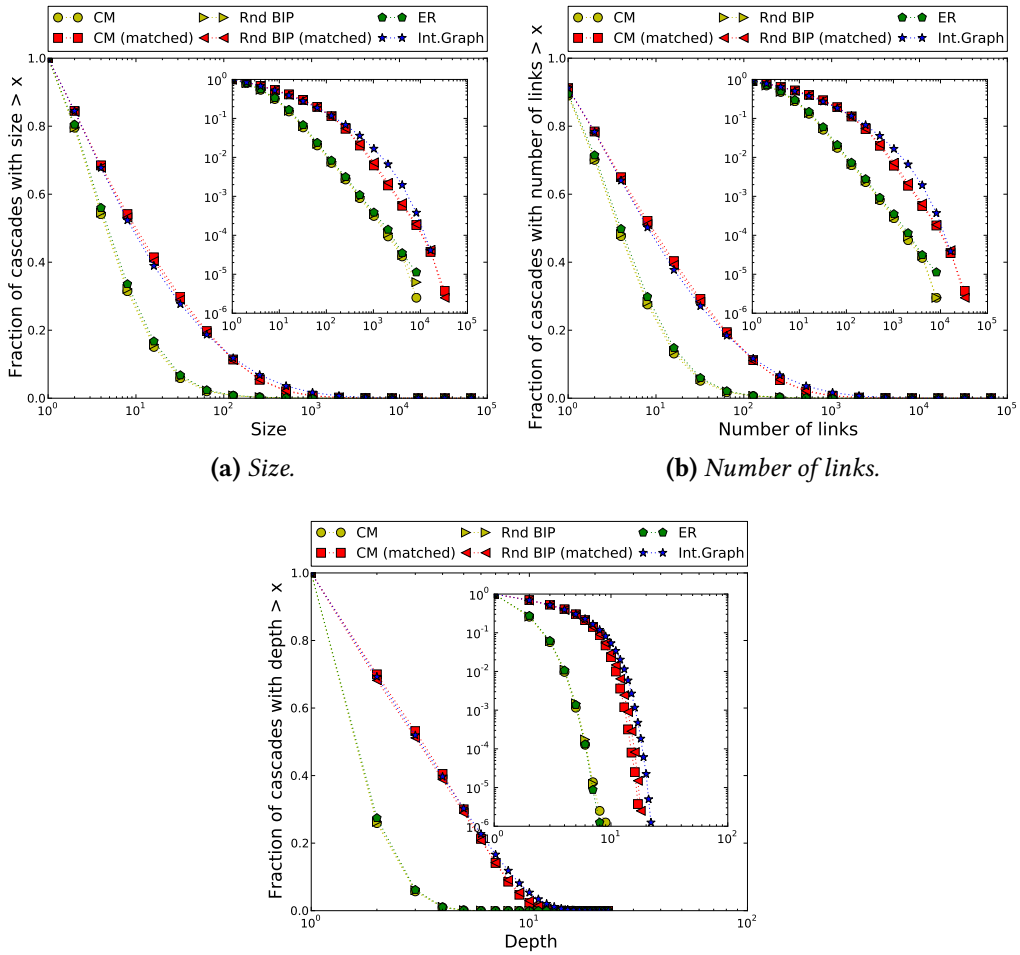
(a) *Size.*

(b) *Number of links.*

(c) *Depth.*

Figure 5.5 – *Complementary cumulative distribution of cascade properties in dynamic random graphs and homogeneous node behavior. Plots shown in lin-log and log-log scale (inset). Spreading model is insensible to variations on network topology and initial providers, with minor variations. All curves in the inset graph are superposed.*

yield sets of cascades with similar profile. This heterogeneous SI model with ICT is sensitive enough to highlight a topological difference between the variations considered, namely the degree distribution of the seeds. Additionally, it shows that in the context of time-bounded simulations like ours, the impact of local clustering might be negligible compared to the degree distribution.



(a) *Size.*

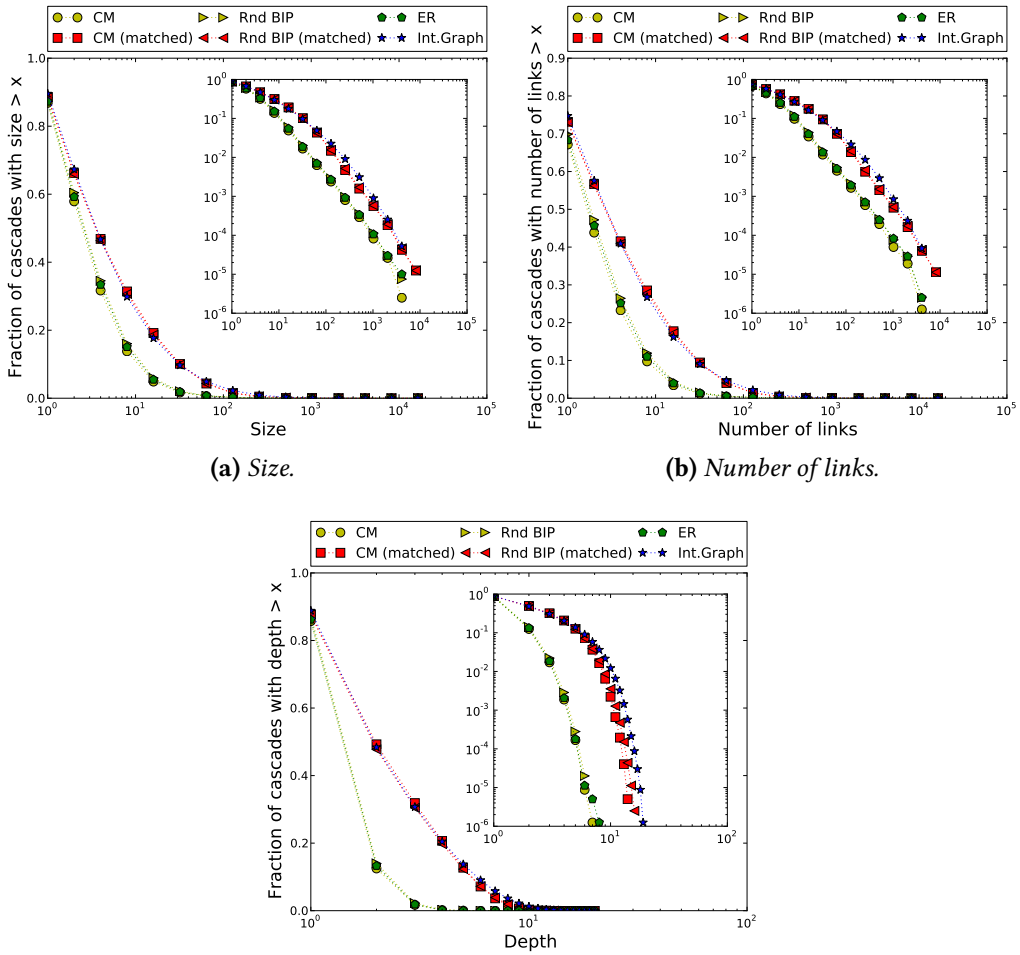
(b) *Number of links.*

(c) *Depth.*

Figure 5.6 – *Complementary cumulative distribution of cascade properties in static random graphs and heterogeneous node behavior. Plots shown in lin-log and log-log scale (inset). Spreading model features the essentially same behavior on all graphs if the degree distribution of initial providers is shuffled and likewise a similar behavior on graphs with similar initial provider degree distribution.*

5.4 Summary

We have inspected the interplay between the underlying network and the model simulating file spreading in different networks. In particular, we have simulated the simple SIR model in a sequence of uniformly random graphs derived from the random graph, with increasing complexity. Furthermore, in terms of the studied properties, the simple SIR model generates similar cascades on random networks having the same degree distribution as the interest graph. We have also found that



(a) *Size.*

(b) *Number of links.*

(c) *Depth.*

Figure 5.7 – *Complementary cumulative distribution of cascade properties in dynamic random graphs and heterogeneous node behavior. Plots shown in lin-log and log-log scale (inset). Spreading model features the essentially same behavior on all graphs if the degree distribution of initial providers is shuffled and likewise a similar behavior on graphs with similar initial provider degree distribution.*

in our setting (with simulation time constraints) the addition of clustering on the random graph did not change the properties of the spreading cascades qualitatively.

Given the improvement in performance brought about integrating temporal patterns into the models, we have tested the impact of the underlying network structure in the diffusion process, using the framework introduced in the previous chapter. More precisely, we have simulated this model on a series of increasingly

realistic random graphs derived from the interest graph. In this case, as throughout the chapter, our simulations have a time-span constraint, to fit the observed time window, in contrast with the usual asymptotic analysis found in the literature. In this "out of the equilibrium" regime, we found that models with homogeneous peer behavior are essentially insensible to all the canonical properties examined. In contrast, models featuring heterogeneous peer behavior were sensible to selected topological properties. More precisely, the topological property with the biggest impact on the simulated cascades was the initial providers degree distribution. Moreover, common graph properties with a demonstrated impact in asymptotic analysis have a minor impact in our simulated cascades. This finding highlights the importance of a frequently overlooked albeit important property in spreading cascade simulations. Also, this result reinforces the rationale to examine models featuring heterogeneous peer behavior.

Chapter 6

Contributions and perspectives

Contents

In this thesis we set out to study quantitatively real-world diffusion, focusing particularly on spreading cascades as our central object of study. The importance of this topologically rich object emerged in recent years, with the advent of several empirical works examining on-line diffusion. Though these works have undoubtedly advanced our knowledge of spreading dynamics, we barely scratched the surface. On the empirical side, it has proven challenging to characterize cascade structure in terms of simple measures, as they generally feature a complex structure. In particular, various cascade properties have been investigated, but to this day there is no consensus on which properties make a satisfactory synthesis of the cascade structure. On the theoretical side, studies in the literature have focused on predicting the fraction of infected individuals in a given population in the long-term – as discussed previously, important asymptotic results relating topological properties of the underlying network (notably in terms of degree distribution and spectrum) were established in the last decade. However, similar theoretical studies in terms of cascade structure theoretical are still an open research area. Hence, a

better understanding of the empirical data, the theoretical models and, particularly, the link between both is also crucial to the characterization of information diffusion in large real-world networks.

We have decided to examine in this thesis the most popular family of network diffusion models, comparing it to real-world diffusion data. In particular, instead of exploring the parameter space of these models in the search of parameters capable of generating realistic cascades, we have investigated if these models generated realistic cascades with the most likely parameter values given our data. That is, we have supposed those models were able to account for the dynamic of the observed diffusion and calibrated the models accordingly, using standard parameter inference techniques. We then compared simulations of the calibrated model to the real-world data. In the following, we discuss this distinctive data-driven approach.

6.1 Summary and contributions

In this section we summarize the contributions of this thesis in context and discuss the challenges we have faced and the the decisions we have made in the course of this work.

6.1.1 Framework and empirical characterization

Our first contribution was to identify a rich dataset for the study of diffusion and propose a framework to do so. As we have discussed in detail in Chapter 3, standard diffusion models are based upon local transmission rules, which take into account the structure of the underlying graph. Therefore, in order to calibrate the spreading model parameters we needed both the underlying graph and the set of spreading cascades. We reconstructed the directed acyclic graphs representing the spreading cascades from the spreading trace. To obtain the interest graph of peers, we proposed a methodology to reconstruct it from the bipartite graph of peers and shared files.

In terms of empirical exploration, we have characterized the spreading cascades in terms of three key structural properties – size, depth and number of links. Standard topological properties of the interest graph were analyzed and we have observed a small diameter, an heterogeneous node degree distribution, low global density and high local clustering. Hence, we have shown that the interest graph topological properties are consistent with the empirical literature on complex networks, and thus, suited for our analysis of diffusion on complex networks. As discussed in Chapter 1, publicly available datasets suffered with missing data regarding the diffusion trace or the underlying network until recently. Though, since the beginning of this thesis a number of large-scale rich datasets have been published, it made sense to gather our own dataset [1].

6.1.2 Inadequacy of the simple SIR and extensions

Turning to the question of the model examination, we have decided to focus our analysis on the most popular family of epidemic diffusion models: the SIR model adapted to networks and derivatives (in particular we begin our analysis with the simplest SIR model). We have compared the real-world data with simulations from model in question, set up to behave as closely as possible to the real file spreading if we assume the file spreading followed the model dynamic. Indeed we have calibrated the parameters with parameter which maximize the likelihood, in agreement with the framework discussed in [Goyal et al., 2010]. In addition to the spreading parameters of the model, we have also identified the initial providers or "seeds" in our dataset to use them as a simulation input. As we remark in the first chapter, despite the numerous papers dealing with the theoretic/asymptotic analysis of these models or their applications, there are surprisingly few papers devoted to the calibration of such network diffusion models with real-world data. Moreover, the question of time bounds in the observed data and its potential impacts is hardly discussed, even though data gathering is frequently bounded in time.

An important methodological challenge in the comparison of simulated and real spreading traces concerned the time bounds of the real-world data: the evolution

1. The dataset description and the empirical findings are summarized in [Bernardes et al., 2012]

of the simple SIR model is given in terms of an intrinsic time (namely, the number of steps in the simulation algorithm), which is not comparable with the time scale of the real diffusion events, as recorded in the trace (measured in "real" time, e.g., seconds). Thus, in order to compare the simulated and real spreading cascades, we have decided to hold one property constant, say size (or depth) and, for each file, generate a simulated cascade with the same size (or depth) as the corresponding real cascade and compare the remaining properties. In other words, let the SIR model spreading parameters be calibrated and a set of seeds for each file be given. If we generate a set of cascades in which each file has the same size (or depth) as the corresponding real cascade, how does the distribution of the other properties for the set of generated cascades compare with the real ones? The simulation results revealed that the simulated cascades were qualitatively different from the real ones. Indeed, real-world cascades were typically more "elongated" and with a greater number of links compared to generated cascades. This finding naturally raised an alarm against the common assumption that diffusion phenomena closely resemble simple epidemic models.

In fairness, the fact that the simple SIR model was unable to generate realistic cascades in the framework considered does not imply that this model is invaluable. Indeed, it is based upon few and simple assumptions, but enough to yield an interesting dynamic. This is positive in and of itself and sufficient to be a potential archetype for the observed diffusion phenomenon. That said, before we conducted the experiment we thought this model would likely be too simple to capture the observed spreading structure so, in this sense, the divergent results were expected. However, we also expected that two natural SIR model extensions, which take into account heterogeneities found in our data (namely file popularity and peer behavior), might generate substantially more realistic cascades. Surprisingly enough, subsequent experiments with these model extensions yielded cascades which remained substantially divergent from the real-world cascades (they were as divergent as the cascades generated with the simple SIR model). In sum, the simplest SIR and two considered extensions (natural as they were, given the data) were insufficient to generate realistic spreading cascades.

At this point it became clear that in order to improve the model, we had to

explore other refinements. Going back to the data, we had observed in Chapter 2 that there is a substantial number of nodes with small degree, which have participated in the P2P system shortly and which have exchanged files with a small number of nodes (which typically feature a high degree). From the perspective of an infecting peer – i.e., a node which has just become infected and is about to infect its neighbors – the probability of infecting any of its neighbors is homogeneous in all the models considered until that point [2]. Therefore, star-like nodes in the graph – i.e., highly connected nodes having a lot of small degree "satellites" connected to it – might contribute to the generation of cascades with bigger size-to-depth ratio. In other words, the interplay of the examined models and the underlying graph might generate cascades less elongated than the ones we observe in the data. Hence, we hypothesized that the missing ingredient in the spreading models examined up to that point might have been a notion of affinity between infecting nodes and their target, which would influence the infection probability between this pair of nodes.

We proposed a straightforward measure of affinity between each pair of peers in the context of P2P file spreading, namely the *interest affinity*, given by the number of files both peers have been interested in. With this extra information the original interest graph becomes a weighted interest graph. Next, we adapted the diffusion dynamics to the weighted graph, assuming that files spread easier between nodes with greater affinity in the spirit of [Onnela et al., 2007], a study of real-world diffusion of information on weighted graphs. Hence, the spreading probability in the adapted diffusion process became a function not only on the infected node, but also in its target through the affinity measure. In other words, it depended on the weight of the edge connecting both nodes. Once these modifications were made, we have calibrated the new model and generated a set of spreading cascades. The new simulated cascades revealed a persistent divergence in cascade shape pattern found previously: they are also typically much shorter and wider compared to the real spreading cascades. Since the impact of the introduction of the affinity measure was qualitatively insignificant we concluded that the absence of this

2. Recall that in the *heterogeneous* models examined in Chapter 3, the probability of infecting one's neighbors might change according to the infecting node itself or according to file being spread, but not according to the infected node's neighbor, as each infecting node does not distinguish its neighbors in its contagion attempts.

parameter was not the primary shortcoming of the original model.

To sum up, we recall that we are interested in evaluating the pertinence of epidemic contagion models to reproduce key structural properties of real-world spreading cascades. We began examining a simple and arguably the most popular network diffusion model in the literature and established its inadequacy to generate realistic spreading cascades, in terms of the patterns found in file spreading cascades on P2P systems. Given the flexibility/generality of the spreading dynamic and the multitude of factors which may have an impact on the diffusion dynamic, it is hard to categorically reject the SIR model as inadequate diffusion model in practice, so we decided to investigate natural extensions to the model, which explore key properties found in the data. We have examined improvements both in terms of the spreading dynamic (heterogeneous models according to peer behavior or file popularity) and in terms of the underlying network structure (interest affinity measure) and found they did not bring about, separately, major changes in the shape of the simulated cascades. These results combined suggest that the the assumption that the simple SIR model models the dynamic of the spreading process, particularly in the context of P2P systems, may not be verified.

6.1.3 Temporal patterns analysis and integration

Although the introduction of the affinity measure did not improve the simple SIR model sensibly, we had an intuition this distinguishing the interaction of peer sharing occasional files and more present peers was a key element missing in the model. Hence we thought about integrating the interaction time directly into the model, namely transforming the original interest graph into a dynamic graph. In this way, the spreading impact of transient nodes would be significantly diminished compared to more present nodes, with a more steady presence in the network. Evidently this significant change in the interest graph presupposes, first, the connection times of each node and an adapted spreading model which would evolve in seconds – that is in "real" time, as opposed to an intrinsic simulation time. Indeed, the spreading process is supposed to interact with the graph, taking into account the nodes and links present in the system at time $t > 0$ measured in

seconds and the process is supposed to evolve in the same time scale.

As we discussed in Chapter 4, although our dataset features temporal data in terms of time stamps for request events, but not the connection events for all peers, we had to reconstruct connection times from the data we had. Using statistical methods we have inferred likely connection and disconnection times for each node in the graph, which in turn we used to reconstruct the dynamic interest graph.

In terms of the diffusion process, we decided to abandon the simple SIR and use instead a SI model with a latency between the time a node becomes infected and the time it infects each of its neighbors, namely the "inter-contagion time" (ICT). Again making a fairly standard assumption that these times are also distributed according to exponential times, we were able to calibrate them using the available data and embed the process with a time scale evolution in terms of seconds, as we wanted. Furthermore, we decided to examine two variants of these models: one in which the ICTs follows the same distribution for all nodes, that is, the node behavior is homogeneous and another in which each node has his own exponential distribution, to account for the heterogeneous behavior of peers. Finally, we adapted the calibrating methods to this new model as we did throughout the model examination.

Once we adapted all parts of the framework we have simulated the temporal SI models on the dynamic interest graph: in terms of cascade size, the results were strikingly improved with respect to the previous simulations. Indeed, we reproduce a set of cascades with similar size as the real-world distribution. The distribution of the number of links was also improved, to a lesser extent. The sole property we could not improve qualitatively was the depth of the simulated cascades, which remained small compared to the real ones. Since we changed two major factors with respect to previous experiments (underlying network and diffusion process), we decided to assess the individual impact of each improvement. Since the dynamic graph presupposes a temporal diffusion process, but not the converse, we decided to simulate the same model on the original (i.e., static) graph and compare with the simulation of the same model in the dynamic graph. Comparing both simulations we conclude that the change in the model alone did not

bring about the improvement in simulated cascade properties found previously. Hence, we conclude that the key improvement is due to the dynamic graph or the combination of the dynamic graph associated with the temporal SI model.

6.1.4 Impact of the underlying network structure

In our quest to identify the relevant factors taken into account by the diffusion model, we have developed an experiment to investigate the impact of the underlying network structure on the simulated spreading cascades. As we have mentioned in the first chapter, there are theoretical results in the literature linking graph properties and the asymptotic behavior of epidemic spreading models such as the SIR model family – summarized for instance in [Barrat et al., 2008]. In this sense we expected a priori that the empirical properties of the underlying network would play a role in the spreading simulation, even though our framework of simulation is outside the scope of these theorems. Indeed, the originality of our approach is precisely analyzing diffusion models in more realistic settings, comparing with real datasets, which are naturally bounded in time. Hence, the interest to uncover the impact of these properties taking into account time constraints.

The interest graph has a rich topological structure, which evolved organically through the interaction of peers sharing files. As pointed out in Chapter 2, it also features properties common to other complex networks, particularly low density, heterogeneous node degree distribution, and local clustering. Evidently these properties are not independent from one another, so in order to assess the impact of these properties individually, we decided to generate a sequence of random graphs, beginning with a baseline graph derived from the interest graph and incrementally adding the properties in question; monitoring the behavior of the simulations from one graph in the series to the other one can identify the impact of each property. Fortunately, in the last decade, methods have been developed to generate uniform random graphs closely matching the target properties mentioned.

We began this analysis using the simple SIR as we describe in Chapter 3. Using the model input computed in our framework (that is, same seeds, spreading parameters and bounds in time) we have simulated this model on all graphs and obtained

essentially the same results for all graphs except the baseline graph. That is, all graphs having same degree distribution yielded the same cascade profiles, which suggest that this graph property had a primary impact on the simulation and that the other properties were unimportant. Next we have performed the same analysis with the temporal SI models (i.e., featuring an inter-contagion time), with two extra variations to measure the impact of the connection patterns and of the seeds' degree distribution. Again using the model inputs obtained in our corresponding framework we have concluded that temporal SI model with homogeneous peer behavior is insensitive to the increment of complex networks' topology properties. In other words, simulated cascades feature the same profile, despite the increment in complex topology properties given by the random graph sequence. In contrast, examining the temporal SI model with heterogeneous peer behavior we found that this model is highly sensitive to the seeds' degree distribution. Indeed, this property was the single most important factor in this case; the other properties of the graph were secondary or unimportant. Given that this temporal SI model was the most realistic model tested, this draws attention to a relevant though overlooked condition for epidemic diffusion models in time-bounded simulations: the seed nodes degree distribution.

6.2 Perspectives

The analysis done in this thesis opens numerous perspectives, which we present in the following. We have grouped them in terms of empirical works and modeling (diffusion and general framework), given that the work presented here is in the intersection of those two domains.

6.2.1 Empirical

On the empirical side, our analysis was founded on the analysis of spreading cascades in terms of their structural profile, characterized by their size, depth and number of links. Though these measures provided a valuable information and made for a rich analysis, they remain very simple compared to the spreading

cascade as an object, a directed acyclic graph. Indeed, it would be interesting to explore other measures which capture overlooked aspects in our analysis (such as motif frequency, cascade clustering, spectrum, etc) or which better represent the spreading cascade. In this sense, [Goel et al., 2012] have proposed a new measure for information cascades, the *Weiner coefficient*, to quantify the virality of the cascade. It would be interesting to characterize the observed cascades in terms of these measures and integrate them to our framework.

Also in terms of empirical approaches, given that most cascades are trivial or quite small, an interesting strategy to better understand the diffusion mechanisms to focus on the rare but most interesting cascades featuring a reasonable number of nodes. Also, it would make clearer the correlation study among different spreading cascade properties. Such move could facilitate the identification of more relevant patterns, potentially in conjunction with the new measures in the previous paragraph, and simplify the identification of more pertinent (albeit specialized) characteristics.

Finally, a major perspective concerning this analysis would be to apply it to other datasets and compare to the results obtained here. Evidently, we acknowledge that many of the difficulties in modeling information spreading we have faced can be the result of data specificity. However, in our defense, it has been argued in [Leibnitz et al., 2006] that spreading of files in P2P follows a SIR-like dynamic: though this claim was purely theoretical and not data-driven, it was still a good additional reason (in addition to the intrinsic qualities of the model and its widespread use) to make a throughout examination of this model and its extensions. In the case the framework yields different assessments of the same model to different data, it will be no doubt interesting to identify the relevant characteristics which justify the difference. If, on the contrary, epidemic models remain unsuitable to other datasets, it will make stronger the case against the careless use of these models when dealing with real-world data.

6.2.2 Diffusion model

In terms of diffusion models, one of the most direct perspectives is to adapt some of the simple SIR model extensions to the temporal SI model and examine the impact of those. Indeed, once established a major factor impacting cascade profiles, one can test the impact of other pertinent factor, which had a second-order impact previously. Namely, it would be interesting to verify the impact of different file popularity in the spreading behavior. The same is true for the weighted graph. Indeed, it would be interesting to assess the impact of the affinity measure in the dynamic graph, since it is possible to combine the weighted graph defined in Chapter 3 with the node connection data to generate a weighted dynamic graph. In terms of weight, we could also try other weight functions, such as considering that very popular files contribute little to the affinity score of a pair of peers, since numerous users possess the file in question; in this sense, rarer files provide more information about the true affinity between two peers. It is hard to guess a priori if this would be a better affinity measure, but no doubt it is worth investigating the impact of other weighted graphs.

Another perspective consists in incorporating tools and results from related fields. In this regard, we have already benefited from a collaboration with colleagues from Université Catholique de Louvain (UCL): we have submitted a joint paper where we proposed a Markovian model to mimic a non-trivial property in the file request profiles in P2P systems (more precisely this study exploited the same dataset we described in this thesis). In particular, our method generates artificial requests trace, similar to our dataset, which can be analyzed as synthetic datasets. We decided to perform the same analysis we have done throughout the thesis to a generated diffusion trace. We have shown that although the file request patterns of peers and the diffusion models are related, the link between the two remains uncovered. In fact, although the artificial trace in question reproduces some realistic request patterns, the corresponding cascades are also qualitatively different from the cascades observed in practice in the real-world data. The analysis which remains to be done is to assess the exact correlation between the two properties, namely spreading cascades and file request patterns.

Finally, an important research perspective is to re-examine the fundamental spreading assumption underlying all the models examined so far. As we have discussed previously, we have examined the most widespread used family of models, the epidemic-inspired SIR model and variations, which assume that the spread of information from one node to the other depends on the influence of the spreader or on the receptivity of the receiver (or on a combination of both). This is not the only possible diffusion mechanism available in the literature. Indeed, another important class of models, threshold or adoption models, assumes that the spread of information depends primarily on the social circle of the receiver. In network terms, the neighboring nodes of the potential receiver node play a key role in the likelihood that he or she "adopts" the information. In this case, nodes are typically more likely to adopt an information if there is a large number (or fraction) of their neighbors which have the information already.

As we mentioned in the first chapter, perhaps the most famous paper associated with this model is [Granovetter, 1978], a sociological study of crowd behavior. In the context of network diffusion, these models were popularized, among others, by Dodd and Watts [Dodds and Watts, 2005]. These works consisted of numerical experiments which explored different scenarios assuming this spreading mechanism. However, despite the interest this approach attracted, very few papers apply it in conjunction with a parameter estimation framework to study of real world cascades as it was done for SIR models in this thesis and elsewhere [Saito et al., 2008, Goyal et al., 2010].

One of these infrequent works, which models the diffusion with an adoption model whose parameters are calibrated using real-world data, is [Bakshy et al., 2009]. The model they use is not the typical model popularized by Dodds and Watts, where the node's neighborhood directly affects the adoption outcome. Rather, they consider a continuous-time model of adoption with stochastic rates of adoption instead of adoption probabilities. The dynamics of this model is as follows: a node enters into state k at the moment that their kth neighbor adopts the information being spread. The model assumes that once an individual is in state k, the time until they adopt, T_k, is exponentially distributed, i.e. they draw an exponentially distributed random variable T_k with mean $1/\lambda_k$ where λ_k will be referred to as the

adoption rate for state k. If a node state changes before they reach their adoption time, they discard that time and draw a new time from the next exponential distribution corresponding to their new state. If one of their existing neighbor adopts, they advance to state $k + 1$.

This model admits extensions that take into account various heterogeneities, is also compatible with dynamic graphs and its parameters can be estimated using a maximum of likelihood. Hence, it would be no doubt interesting to develop the same study conducted in this thesis using this alternative diffusion model instead of the epidemic models suggested so far. As we mentioned in the Empirical perspectives section above, theoretical works on information diffusion models in the context of P2P file sharing systems wagered on epidemic models, particularly the SIR model, as the best candidate to describe information diffusion in these networks, so it made sense to investigate these models exhaustively with priority.

6.2.3 General

In a more theoretical note, the first general perspective opened with this study, particularly as a consequence of the study of the network topology impact, is to extend the analysis using other model inputs. As discussed previously, the set of seeds, spreading parameters and time bounds were determined as a function of our framework, which in turn dealt with a real dataset. Ideally, for each input parameter it would be interesting to vary the values considered to have a better sense how these important diffusion models behave in constrained time. This is both an important and straightforward perspective, given what has already been developed in this thesis.

Still regarding the question of better understanding the behavior of these models in particular settings, examining other kinds of underlying networks would be interesting, particularly other kinds of dynamic graphs. The dynamic graphs we have studied consisted essentially of static graphs made dynamic following the connection and disconnection of peers; thus, if any two peers remain on-line, no changes in the corresponding link between the two nodes will change. In

contrast, other kinds of dynamic graphs, such as contact networks, are such that the evolution of graphs is usually given by the appearance and disappearance of links between individuals. This different "evolution nature" would likely impact the models differently. Indeed, we have seen recently a growing interest in the interplay between epidemic contagion models and this kind of dynamic graph in the literature from a theoretical angle [Karimi and Holme, 2013, Lambiotte et al., 2013].

In conclusion, once we have accumulated sufficient information on the evolution of these models in all the aspects listed so far, we will be able to identify a representative enough behavior of these models to develop a direct way to test the hypothesis that a certain empirical diffusion trace can be explain by such models. In this sense, a Bayesian statistics approach may prove interesting, as the evaluation process involves calibrating the model with the most realistic parameters given the data and we might have some a priori knowledge of the parameters.

In any case, many research directions remain open and, although characterizing information on complex networks is no doubt challenging, the relevance of this subject remains great, from a purely scientific perspective as well as from an applied perspective.

Bibliography

[Adamic, 1999] Adamic, L. A. (1999). The small world web. In *Research and Advanced Technology for Digital Libraries*, pages 443–452. Springer. 19

[Adamic and Huberman, 2001] Adamic, L. A. and Huberman, B. A. (2001). The web's hidden order. *Commun. ACM*, 44(9):55–59. 19

[Adar and Adamic, 2005] Adar, E. and Adamic, L. A. (2005). Tracking information epidemics in blogspace. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 207–214. IEEE. 22

[Adar and Huberman, 2000] Adar, E. and Huberman, B. (2000). Free riding on gnutella. *First Monday*, 5(10-2). 31

[Aidouni et al., 2009] Aidouni, F., Latapy, M., and Magnien, C. (2009). Ten weeks in the life of an edonkey server. In *23rd IEEE International Symposium on Parallel and Distributed Processing, IPDPS 2009, Rome, Italy, May 23-29, 2009*, pages 1–5. 29

[Anderson and May, 1991] Anderson, R. and May, R. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Science Publications, Oxford. 20

[Andersson and Britton, 2000] Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis (Lecture Notes in Statistics) (v. 151)*. Springer, 1 edition. 20

[Bakshy et al., 2011] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). 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. 22, 27

[Bakshy et al., 2009] Bakshy, E., Karrer, B., and Adamic, L. A. (2009). Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 325–334, New York, NY, USA. ACM. 22, 96

[Ball, 2011] Ball, P. (2011). News mining might have predicted arab spring. *Nature News*. 13

[Ban et al., 2011] Ban, T., Guo, S., Zhang, Z., Ando, R., and Kadobayashi, Y. (2011). Practical network traffic analysis in p2p environment. In *Proceedings of the 7th International Conference on Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 1801–1807. IEEE. 28

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512. 19

[Barabási and Oltvai, 2004] Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113. 17

[Barrat et al., 2008] Barrat, A., Barthlemy, M., and Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge U. Press, New York, NY, USA. 92

[Bass, 1969] Bass, F. (1969). A new product growth for model consumer durables. *Management Sciences*, 15(1):215–227. 21

[Bernardes et al., 2012] Bernardes, D. F., Latapy, M., and Tarissan, F. (2012). Relevance of sir model for real-world spreading phenomena: Experiments on a large-scale p2p system. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. Istanbul, Turkey; 2012-08-26 – 2012-08-29. 36, 87

[Bernardes et al., 2013] Bernardes, D. F., Latapy, M., and Tarissan, F. (2013). Inadequacy of sir model to reproduce key properties of real-world spreading cascades: experiments on a large-scale p2p system. *Social Network Analysis and Mining*, 3(4):1195–1208.

[Bollobás, 1998] Bollobás, B. (1998). *Modern graph theory*, volume 184. Springer. 16

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703. 19

[Coupechoux and Lelarge, 2012] Coupechoux, E. and Lelarge, M. (2012). How clustering affects epidemics in random networks. *CoRR*, abs/1202.4974. 71

[Crandall et al., 2008] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., and Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM. 16

[de Tarde and Parsons, 1903] de Tarde, G. and Parsons, E. (1903). *The Laws of Imitation*. H. Holt. 15

[Diestel, 2010] Diestel, R. (2010). *Graph Theory*. Graduate Texts in Mathematics, Volume 173. Springer-Verlag, Heidelberg, 4th edition. 16, 38

[Doan et al., 2012] Doan, S., Vo, B.-K., and Collier, N. (2012). An analysis of twitter messages in the 2011 tohoku earthquake. In Kostkova, P., Szomszor, M., and Fowler, D., editors, *Electronic Healthcare*, volume 91 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 58–66. Springer Berlin Heidelberg. 13

[Dodds and Watts, 2005] Dodds, P. S. and Watts, D. J. (2005). A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–604. 22, 96

[Dow et al., 2013] Dow, P. A., Adamic, L. A., and Friggeri, A. (2013). The anatomy of large facebook cascades. In Kiciman, E., Ellison, N. B., Hogan, B., Resnick, P., and Soboroff, I., editors, *ICWSM*. The AAAI Press. 27

[Dunne, 2006] Dunne, J. A. (2006). The network structure of food webs. *Ecological networks: linking structure to dynamics in food webs*, pages 27–86. 17

[Easley and Kleinberg, 2010] Easley, D. A. and Kleinberg, J. M. (2010). *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press. 16

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262. 19

[Fertik and Thompson, 2010] Fertik, M. and Thompson, D. (2010). *Wild West 2.0: How to Protect and Restore Your Reputation on the Untamed Social Frontier*. AMACOM. 13

[Fortunato, 2010] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174. 19

[Foss et al., 2013] Foss, S., Korshunov, D., and Zachary, S. (2013). *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer Series in Operations Research and Financial Engineering. Springer. 18

[Freeman, 2004] Freeman, L. C. (2004). *The development of social network analysis*. Empirical Press Vancouver. 14

[Friggeri et al., 2011] Friggeri, A., Cointet, J.-P., Latapy, M., et al. (2011). A real-world spreading experiment in the blogosphere. *Complex Systems*, 19(3):235. 15, 23

[Goel et al., 2013] Goel, S., Anderson, A., Hofman, J., and Watts, D. (2013). The structural virality of online diffusion. *Preprint*. 24, 27

[Goel et al., 2012] Goel, S., Watts, D. J., and Goldstein, D. G. (2012). The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 623–638. ACM. 22, 41, 93

[Gomez-Rodriguez et al., 2012] Gomez-Rodriguez, M., Leskovec, J., and Krause, A. (2012). Inferring networks of diffusion and influence. *ACM Trans. Knowl. Discov. Data*, 5(4):21:1–21:37. 24, 36

[Goyal et al., 2010] Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM. 24, 87, 96, 111

[Granovetter, 1978] Granovetter, M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443. 14, 22, 96

[Guha et al., 2004] Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. (2004). Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412. ACM. 17

[Guillaume and Latapy, 2004] Guillaume, J.-L. and Latapy, M. (2004). Bipartite structure of all complex networks. *Inf. Process. Lett.*, 90(5):215–221. 72

[Gurcel and Watier, 2002] Gurcel, L. D. and Watier, P. (2002). *La Sociologie de Georg Simmel (1908). Éléments actuels de modélisation sociale*. Presses Universitaires de France, Paris. 19

[Handurukande et al., 2006] Handurukande, S. B., Kermarrec, A.-M., Le Fessant, F., Massoulié, L., and Patarin, S. (2006). Peer sharing behaviour in the edonkey network, and implications for the design of server-less file sharing systems. In *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*, EuroSys '06, pages 359–371, New York, NY, USA. ACM. 32, 36, 41

[Hayek, 1948] Hayek, F. (1948). *Individualism and Economic Order*. Midway reprints. University of Chicago Press. 14

[Huberman and Adamic, 1999] Huberman, B. A. and Adamic, L. A. (1999). Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131. 17

[Hughes et al., 2005] Hughes, D., Coulson, G., and Walkerdine, J. (2005). Free riding on gnutella revisited: the bell tolls? *Distributed Systems Online, IEEE*, 6(6). 32

[Huxley and Huxley, 1947] Huxley, T. and Huxley, J. (1947). *Evolution and ethics: 1893-1943*. Pilot Press. 14

[Iamnitchi et al., 2011] Iamnitchi, A., Ripeanu, M., Santos-Neto, E., and Foster, I. (2011). The small world of file sharing. *IEEE Trans. Parallel Distrib. Syst.*, 22(7):1120–1134. 36

[Isella et al., 2011] Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and Van den Broeck, W. (2011). What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180. 17

[Jackson and Rogers, 2005] Jackson, M. O. and Rogers, B. W. (2005). The economics of small worlds. *Journal of the European Economic Association*, 3(2-3):617–627. 19

[Jewell, 1982] Jewell, N. (1982). Mixtures of exponential distributions. *The Annals of Statistics*, pages 479–484. 61

[Karimi and Holme, 2013] Karimi, F. and Holme, P. (2013). A temporal network version of watts's cascade model. In *Temporal Networks*, pages 315–329. Springer. 97

[Kempe and Kleinberg, 2002] Kempe, D. and Kleinberg, J. (2002). Protocols and impossibility results for gossip-based communication mechanisms. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 471–480. IEEE. 22

[Kermack and McKendrick, 1927] Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society, London*, 115:700+. 15, 20

[Kittur and Kraut, 2008] Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM. 16

[Klein, 2012] Klein, D. (2012). *Knowledge and Coordination: A Liberal Interpretation*. Oxford University Press, USA. 14

[Kleinberg, 2006] Kleinberg, J. (2006). Complex networks and decentralized search algorithms. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 1019–1044. 18

[Kleinberg et al., 1999] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: Measurements, models, and methods. In *Computing and combinatorics*, pages 1–17. Springer. 17, 19

[Kolaczyk, 2009] Kolaczyk, E. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer. 19

[Kossinets and Watts, 2006] Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90. 17

[Kulbak et al., 2005] Kulbak, Y., Bickson, D., et al. (2005). The emule protocol specification. *eMule project, http://sourceforge. net*. 28

[Kumar et al., 2004] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39. 17

[Lambiotte et al., 2013] Lambiotte, R., Tabourier, L., and Delvenne, J.-C. (2013). Burstiness and spreading on temporal networks. *The European Physical Journal B*, 86(7):1–4. 97

[Latapy et al., 2013] Latapy, M., Magnien, C., and Fournier, R. (2013). Quantifying paedophile activity in a large p2p system. *Information Processing and Management*, 49(1):248 – 263. 13

[Latapy et al., 2008] Latapy, M., Magnien, C., and Vecchio, N. D. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31 – 48. 36

[Le Bon, 1895] Le Bon, G. (1895). *The Crowd a Study of the Popular Mind*. Kessinger Publishing. 15

[Leibnitz et al., 2006] Leibnitz, K., Hossfeld, T., Wakamiya, N., and Murata, M. (2006). Modeling of epidemic diffusion in peer-to-peer file-sharing networks. In *Proceedings of the Second international conference on Biologically Inspired Approaches to Advanced Information Technology*, BioADIT'06, pages 322–329, Berlin, Heidelberg. Springer-Verlag. 64, 94

[Leskovec and Horvitz, 2008] Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM. 17, 18, 27

[Leskovec et al., 2007] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. S., and Hurst, M. (2007). Patterns of cascading behavior in large blog graphs. In *SDM*. SIAM. 13, 17, 22, 41

[Leskovec et al., 2006] Leskovec, J., Singh, A., and Kleinberg, J. (2006). Patterns of influence in a recommendation network. In *Advances in Knowledge Discovery and Data Mining*, pages 380–389. Springer. 22

[Lessig, 2002] Lessig, L. (2002). *The Future of Ideas: The Fate of the Commons in a Connected World*. Vintage. Knopf Doubleday Publishing Group. 13

[Liben-Nowell and Kleinberg, 2008] Liben-Nowell, D. and Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638. 23, 41, 56

[Liljeros et al., 2001] Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., and Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908. 19

[Mill, 1843] Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and methods of scientific investigation*. J. W. Parker, London. 14

[Molloy and Reed, 1995] Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180. 72

[Neiger et al., 2012] Neiger, V., Crespelle, C., and Fleury, E. (2012). On the structure of changes in dynamic contact networks. In *SITIS*, pages 731–738. IEEE. 17

[Newman, 2010] Newman, M. (2010). *Networks: An Introduction*. OUP Oxford. 18

[Newman, 2001] Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409. 17, 18, 19

[Newman, 2003] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256. 71, 72

[Onnela et al., 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths

in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336. 17, 22, 55, 89, 113

[Pareto, 1897] Pareto, V. (1896–1897). *Cours d'économie politique professé a l'université de Lausanne. (French) [Course on political economy given at the University of Lausanne]*. F. Rouge, Lausanne, Switzerland. 19

[Pastor-Satorras and Vespignani, 2001] Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200. 22, 71

[Prakash et al., 2012] Prakash, B. A., Chakrabarti, D., Valler, N. C., Faloutsos, M., and Faloutsos, C. (2012). Threshold conditions for arbitrary cascade models on arbitrary networks. *Knowledge and information systems*, 33(3):549–575. 71

[Price, 1976] Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306. 17

[Ritzer, 2007] Ritzer, G. (2007). *Modern Sociological Theory*. McGraw-Hill Higher education. McGraw-Hill Higher Education. 14

[Saito et al., 2008] Saito, K., Nakano, R., and Kimura, M. (2008). Prediction of information diffusion probabilities for independent cascade model. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 67–75. Springer. 24, 96

[Salah Brahim et al., 2011] Salah Brahim, A., Le Grand, B., Tabourier, L., and Latapy, M. (2011). Citations among blogs in a hierarchy of communities: Method and case study. *Journal of Computational Science*, 2(3):247–252. 17

[Saroiu et al., 2002] Saroiu, S., Gummadi, K. P., Dunn, R. J., Gribble, S. D., and Levy, H. M. (2002). An analysis of internet content delivery systems. *ACM SIGOPS Operating Systems Review*, 36(SI):315–327. 32

[Schumpeter, 1909] Schumpeter, J. (1909). On the concept of social value. *The Quarterly Journal of Economics*, 23(2):pp. 213–232. 14

[Sen and Wang, 2004] Sen, S. and Wang, J. (2004). Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Trans. Netw.*, 12(2). 28

[Sharma et al., 2012] Sharma, N. K., Ghosh, S., Benevenuto, F., Ganguly, N., and Gummadi, K. (2012). Inferring who-is-who in the twitter social network. *ACM SIGCOMM Computer Communication Review*, 42(4):533–538. 17

[Smith, 1789] Smith, A. (1789). *An Inquiry Into the Nature and Causes of the Wealth of Nations*. An Inquiry Into the Nature and Causes of the Wealth of Nations. A. Strahan and T. Cadell. 14

[Sporns et al., 2004] Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in cognitive sciences*, 8(9):418–425. 17

[Sun et al., 2009] Sun, E., Rosenn, I., Marlow, C., and Lento, T. M. (2009). Gesundheit! modeling contagion through facebook news feed. In *ICWSM*. 22

[Travers and Milgram, 1969] Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4):425–443. 18

[Tredennick, 1933] Tredennick, H. (1933). *The Metaphysics: Books I to IX, with an English Translation by H. Tredennick*. Loeb classical library. Heinemann. 14

[Wang et al., 2003] Wang, Y., Chakrabarti, D., Wang, C., and Faloutsos, C. (2003). Epidemic spreading in real networks: An eigenvalue viewpoint. In *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on*, pages 25–34. IEEE. 71

[Watts, 2011] Watts, D. (2011). *Everything Is Obvious*. Atlantic Books, Limited. 14

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442. 18, 19

[Wotal et al., 2006] Wotal, B., Green, H., Williams, D., and Contractor, N. (2006). Wow!: The dynamics of knowledge networks in massively multiplayer online role playing games (mmorpg. In *In Sunbelt XXVI: International Sunbelt Social Network Conference*. 16

[Zipf, 1948] Zipf, G. K. (1948). On the number, circulation-sizes, and the probable purchasers of newspapers. *The American Journal of Psychology*, 61(1):79–89. 19

Appendix A

Résumé

Dans cette thèse, nous avons étudié la diffusion de l'information dans les grands *graphes de terrain* (des réseaux d'interaction complexes réels), en se focalisant particulièrement sur les patterns structurels de la propagation. Plus précisément, notre objet d'étude central est la *cascade de diffusion*, c'est-à-dire, le graphe qui relie les nœuds du réseau (qui représentent des individus, machines, etc) par où l'information est passée, en mettant en évidence "qui a transmis l'information à qui". Cet objet topologiquement riche a reçu beaucoup d'attention depuis quelques années grâce à la disponibilité de traces numériques détaillées sur des événements de diffusion en ligne (email, fichiers, tweets, etc.). Sur le plan empirique, il s'est avéré difficile de capturer la structure des cascades de diffusion en termes de mesures simples. Diverses propriétés des cascades ont été étudiées, mais on n'a pas encore trouvé un ensemble de propriétés simples permettant de synthétiser la structure des cascades. Sur le plan théorique, l'approche classique consiste à étudier des modèles stochastiques de contagion et de percolation sur des graphes aléatoires ou réguliers. Le traitement analytique de ce type de modèle sur des structures discrètes s'avère difficile, mais malgré la complexité, plusieurs résultats concernant le comportement asymptotique de modèles simples sont apparus dans la littérature. Néanmoins, le champ d'application de ces modèles reste limité, car les cascades réelles ont généralement une topologie complexe et le processus de diffusion se produit dans une fenêtre de temps limitée (généralement pas assez grande pour l'analyse asymptotique). Par conséquent, une meilleure compréhension des données empiriques, des modèles théoriques et du lien entre les deux est également cruciale pour la caractérisation de la diffusion dans les grands graphes de terrain.

Ce document est organisé de la manière suivante : nous commençons, au premier chapitre, par un état de l'art sur les graphes de terrain et la diffusion dans ce

contexte. Dans le chapitre 2, nous décrivons notre jeu de données et discutons sa pertinence dans le contexte du premier chapitre. Nous présentons la procédure de reconstruction du graphe sous-jacent (où se passe la diffusion) et des cascades de diffusion. Ensuite, dans le chapitre 3, nous évaluons la pertinence d'un des modèles classiques de diffusion sur les réseaux : le modèle SIR. Nous examinons également quelques extensions de ce modèle qui prennent en compte des hétérogénéités de notre jeu de données, ainsi qu'un raffinement du processus de reconstruction du graphe d'intérêt. Dans le chapitre 4, nous explorons la prise en compte du temps dans l'évolution du réseau sous-jacent et dans le modèle de diffusion. Dans le chapitre 5, nous évaluons l'impact de la structure du graphe sous-jacent sur la structure des cascades de diffusion générées avec les modèles étudiés dans les chapitres précédents. Nous terminons ce document par un bilan des résultats (que nous résumons dans la suite) et des perspectives ouvertes par les travaux menés dans cette thèse.

A.0.4 Méthodologie et caractérisation empirique des cascades de diffusion

Notre première contribution a été d'identifier un ensemble de données riche pour l'étude de la diffusion et de proposer une méthodologie d'analyse. Les modèles de diffusion classiques sont basés sur les règles de transmission locales, qui prennent en compte la structure du graphe sous-jacent. Par conséquent, afin de calibrer les paramètres du modèle de diffusion nous avions besoin à la fois du graphe sous-jacent et des cascades de diffusion. Nous avons ainsi construit – dans un premier temps – des graphes acycliques orientés représentant les cascades de diffusion à partir de la trace d'interaction des utilisateurs. Pour obtenir le graphe d'intérêt des pairs, nous avons proposé une méthodologie pour le reconstruire à partir du graphe biparti des pairs et des fichiers partagés.

En termes d'exploration empirique, nous avons caractérisé les cascades de diffusion en termes de trois propriétés structurelles – taille, profondeur et nombre de liens. En analysant des propriétés topologiques standards du graphe d'intérêt des pairs, nous avons observé que le graphe a un petit diamètre, une distribution de

degrés hétérogène, une faible densité globale et un fort *clustering* local. Ainsi, nous avons montré que les propriétés topologiques du graphe d'intérêt sont compatibles avec la littérature empirique sur les graphes de terrain, et donc adaptées à notre analyse, qui se focalise sur la diffusion dans les grands graphes de terrain.

A.0.5 Pertinence du modèle SIR simple et de ses extensions

Quant à la question de l'évaluation du modèle, nous avons décidé de concentrer notre analyse sur la famille la plus populaire de modèles de diffusion inspirées de l'épidémiologie : les modèles SIR. Nous avons comparé les données réelles avec des simulations du modèle SIR simple, calibré en supposant que ce modèle capture bien la dynamique de propagation. En effet, nous avons inféré les paramètres les plus vraisemblables, en accord avec le cadre discuté dans [Goyal et al., 2010]. Nous avons également identifié les fournisseurs originaux ou *graines de la diffusion* dans notre jeu de données pour les utiliser comme une entrée de simulation. Malgré les nombreux articles traitant de l'analyse formelle de ces modèles ou de leurs applications, il y a étonnamment peu d'articles consacrés à l'étalonnage de ces modèles de diffusion de réseau avec des données réelles. En outre, la question des limites de temps dans les données mesurées en pratique et de ses impacts potentiels est à peine abordée, même si la collecte de données est souvent limitée dans le temps.

Un défi méthodologique important dans la comparaison des traces réelles et simulées concerne l'étalement de la durée des simulations : l'évolution du modèle SIR simple est donnée en termes de temps intrinsèque (à savoir, le nombre d'étapes dans l'algorithme de simulation), ce qui n'est pas comparable aux données temporelles sur la trace réelle de diffusion (mesurée en secondes). Ainsi, afin de comparer les cascades de diffusion simulées et réelles, nous avons décidé de tenir une propriété constante, disons taille (ou profondeur) et, pour chaque fichier, générer une cascade simulée avec la même taille (ou profondeur) que la cascade réelle correspondante et comparer les autres propriétés. Autrement dit, étant donnés le modèle SIR calibré et un ensemble de graines de diffusion identifiées pour chaque fichier : si nous générons, pour chaque fichier, des cascades de diffusion

de même taille (ou profondeur) que la cascade réelle correspondante, comment la distribution des autres propriétés de l'ensemble de cascades générées se compare avec les propriétés des vrais cascades ? Les résultats de la simulation ont montré que les cascades simulées sont qualitativement différentes des cascades réelles. En effet, elles sont généralement plus "allongées" et ont un plus grand nombre de liens par rapport aux cascades simulées. Cette constatation remet en cause l'hypothèse courante selon laquelle les phénomènes de diffusion réels ont une dynamique très proche de celle des modèles épidémiques simples.

En toute justice, le fait que le modèle SIR simple a été incapable de générer des cascades réalistes dans le cadre considéré n'implique pas que ce modèle est inintéressant. Au contraire, il est capable de produire une dynamique de contagion non-triviale avec très peu d'hypothèses de base. Cela dit, avant de mener l'expérience nous pensions que ce modèle serait probablement trop simple pour décrire la structure du phénomène de diffusion observé et, dans ce sens, les résultats divergents étaient attendus. Toutefois, nous nous attendions aussi à ce que les extensions du modèle SIR qui prennent en compte la popularité des fichiers et le comportement des pairs pourraient générer des cascades beaucoup plus réalistes que le modèle de base. Contrairement à nos attentes, les expériences avec ces extensions du modèle ont généré des cascades toujours sensiblement divergentes des cascades réelles (et des cascades générées avec le modèle SIR simples). En somme, le modèle SIR simple et les deux extensions qui tiennent compte des hétérogénéités trouvés dans nos données étaient insuffisantes pour générer des cascades de diffusion structurellement réalistes.

Ainsi, pour améliorer le modèle, nous avons dû explorer d'autres possibilités. Pour en revenir aux données, nous avons observé qu'il y a un nombre important de nœuds avec un petit degré, qui ont fait peu d'échanges de fichier, typiquement avec des nœuds ayant un haut degré. Par ailleurs, du point de vue d'un nœud infectant (i.e., un nœud qui vient d'être infecté et est sur le point d'infecter ses voisins) la probabilité d'infecter l'un de ses voisins est homogène dans tous les modèles considérés jusque-là. Par conséquent, les nœuds en forme d'étoile dans le graphe – c'est à dire, les nœuds fortement connectés à des nœuds "satellites" avec petit degré – pourraient contribuer à la génération de cascades avec un plus

grand rapport taille-profondeur. Autrement dit, l'interaction des modèles étudiés et le graphe sous-jacent des pairs peut générer des cascades moins allongées que celles que nous observons dans les données. Nous avons soupçonné ainsi que l'ingrédient manquant dans les modèles de diffusion examinés pouvais être une notion d'affinité entre les nœuds, qui influencerait la probabilité d'infection.

Nous avons, alors, proposé une mesure d'affinité simple entre chaque couple de pairs dans le contexte des échanges de fichiers P2P, à savoir l'affinité en terme des intérêts en commun, donnée par le nombre de fichiers que deux pairs possèdent en commun. Avec cette information supplémentaire le graphe d'intérêt initial devient un graphe pondéré. Ensuite, nous avons adapté la dynamique de diffusion au graphe d'intérêt pondéré, en supposant que les fichiers se diffusent plus facilement entre les nœuds avec une plus grande affinité, comme proposé dans [Onnela et al., 2007]. Par conséquent, la probabilité d'infection du processus de diffusion calibré devient une fonction non seulement du nœud infecté, mais également de sa cible en fonction de la mesure d'affinité. En d'autres termes, il dépend du poids de l'arête reliant les deux nœuds. Une fois ces modifications apportées, nous avons calibré le nouveau modèle et généré un ensemble de cascades simulées. Ces nouvelles cascades ont révélé une divergence persistante : elles sont aussi généralement beaucoup moins profondes et plus larges que les cascades réelles. Étant donné que l'impact de l'introduction de la mesure d'affinité a été qualitativement négligeable, nous avons conclu que l'absence de ce paramètre n'a pas été le principal handicap de la modélisation de base.

Pour résumer, nous nous sommes intéressés à évaluer la capacité des modèles de contagion populaires à reproduire les propriétés structurelles des cascades de diffusion réelles. Nous avons commencé par l'évaluation du modèle le plus populaire dans la littérature et établi son incapacité à générer des cascades de diffusion structurellement réalistes, comparé aux cascades de diffusion de fichiers observées sur les systèmes P2P. Compte tenu de la multitude de facteurs qui peuvent avoir un impact sur la dynamique de diffusion, nous avons décidé d'évaluer des extensions naturelles du modèle SIR, qui explorent les propriétés clés trouvées dans notre jeu de données. Nous avons examiné les améliorations à la fois en termes de dynamique de propagation (modèles hétérogènes selon le comporte-

ment des pairs ou la popularité des fichiers) et en termes de structure du réseau sous-jacent (mesure d'affinité). Nous avons constaté qu'elles n'ont pas apporté des changements majeurs dans la forme des cascades simulées. Ces résultats combinés découragent le choix du modèle SIR pour modéliser la dynamique de diffusion réelle, particulièrement dans le contexte des systèmes de P2P.

A.0.6 Patterns temporaux et leur intégration dans le modèle

Bien que l'introduction de la mesure d'affinité n'ait pas amélioré sensiblement le modèle SIR simple, nous étions persuadés que distinguer l'interaction des utilisateurs occasionnels des utilisateurs plus présents était un élément clé manquant dans le modèle. Ainsi nous avons décidé d'intégrer le temps de présence directement dans la modélisation, en transformant le graphe sous-jacent en un graphe dynamique. De cette manière, l'impact des nœuds peu présents serait diminué de façon significative par rapport à l'impact des nœuds présents plus régulièrement dans le réseau. Évidemment, cette modification importante dans le graphe sous-jacent nécessite, d'une part, des temps de connexion de chaque nœud et un modèle de diffusion adapté qui évolue en secondes – c'est-à-dire, en "temps réel", par opposition à un temps de simulation intrinsèque. En effet, le processus de diffusion est supposé interagir avec le graphe, en tenant compte des nœuds et des liens présents dans le système à tout instant de temps donné (mesuré en secondes).

Notre jeu de données contient des données temporelles (en termes d'horodatage) des requêtes, mais pas des événements de connexion des pairs, donc nous avons dû reconstruire ces temps de connexion à partir des données des requêtes. En utilisant des méthodes statistiques, nous avons inféré les temps de connexion et de déconnexion de chaque pair et nous les avons utilisés pour reconstruire le graphe d'intérêt dynamique.

En terme de processus de diffusion, nous avons décidé d'abandonner le modèle SIR simple et d'utiliser à la place un modèle SI avec une latence entre le moment où un nœud devient infecté et le temps où il infecte chacun de ses voisins, à savoir le "temps inter-contagion" (TIC). En faisant l'hypothèse que ces temps sont exponentiellement distribués, nous avons pu calibrer ces TICs individuels en utilisant

les données disponibles. En outre, nous avons décidé d'examiner deux variantes de ce modèle : l'une dans laquelle le TIC suit la même distribution pour tous les nœuds (comportement homogène) et une autre dans laquelle chaque nœud a sa propre distribution exponentielle, pour tenir compte du comportement hétérogène des pairs. Enfin, nous avons adapté les méthodes de calibrage utilisées jusqu'à présent à ce nouveau modèle.

Une fois la méthodologie adaptée, nous avons simulé le modèle SI temporel sur le graphe d'intérêt dynamique : en termes de taille de cascade, les résultats ont été remarquablement améliorés par rapport aux simulations précédentes. En effet, nous avons généré un ensemble de cascades avec une taille similaire à celle de la distribution réelle. La distribution du nombre de liens a été également améliorée, dans une moindre mesure. La seule propriété que nous n'avons pas pu améliorer qualitativement est la profondeur des cascades simulées, qui reste faible par rapport à celle des cascades réelles. Puisque nous avons changé deux facteurs majeurs par rapport à la modélisation précédente (réseau sous-jacent et processus de diffusion), nous avons décidé d'évaluer l'impact individuel de chaque amélioration. Comme le graphe dynamique suppose un processus de diffusion temporelle, mais pas l'inverse, nous avons décidé de simuler le même modèle sur le graphe d'intérêt original et de le comparer avec la simulation du même modèle dans le graphe dynamique. En comparant les deux simulations, nous concluons que le changement au niveau du modèle exclusivement n'a pas apporté l'amélioration des propriétés de cascade simulées trouvées précédemment. Nous en avons alors déduit que l'amélioration est causée par le graphe dynamique ou par la combinaison du graphe dynamique associé au modèle SI temporel.

A.0.7 Impact de la structure du réseau sous-jacent

Dans notre quête pour identifier les facteurs pertinents pour le choix du modèle de diffusion, nous avons proposé une expérience pour étudier l'impact de la structure du réseau sous-jacent sur la strucure des cascades de diffusion simulées. Des résultats théoriques de la littérature relient les propriétés des graphes et le comportement asymptotique des modèles de diffusion épidémiologiques simples tels

que la famille de modèles SIR. Dans ce sens, nous nous attendions a priori à ce que les propriétés empiriques du réseau sous-jacent jouent un rôle dans la propagation simulée, même si notre cadre de simulation ne satisfait pas les hypothèses de ces théorèmes. En effet, une contribution essentielle de notre travail est précisément d'analyser les modèles de diffusion dans des conditions plus réalistes, en les comparant avec des données réelles, obtenues dans une fenêtre de temps bornée. Dans ce régime de temps borné, nous ne disposons pas d'outils théoriques qui relient les propriétés des graphes et la dynamique du modèle. Même les quelques théorèmes asymptotiques disponibles ne portent pas sur des propriétés structurelles des cascades de diffusion ; ils se concentrent sur la fraction globale de nœuds infectés et sur la probabilité d'extinction de la propagation. Ceci montre l'intérêt de découvrir l'impact de ces propriétés dans notre cadre.

Le graphe d'intérêt a une structure topologique riche, qui a évolué par l'interaction des pairs partageant des fichiers. Il détient des propriétés communes à d'autres graphes de terrain, particulièrement une faible densité globale/fort clustering local et une distribution des degrés hétérogène. Évidemment ces propriétés ne sont pas indépendantes les unes des autres, donc afin d'évaluer l'impact de ces propriétés individuellement, nous avons décidé de produire une séquence de graphes aléatoires, à commencer par un graphe de base (dérivé du graphe d'intérêt) et d'ajouter progressivement les propriétés en question. En surveillant le changement sur la structure des cascades simulées d'un graphe d'une séquence à l'autre on peut identifier l'impact de chaque propriété. Nous avons utilisé des méthodes modernes pour générer des graphes aléatoires uniformes avec les propriétés cibles mentionnées.

Nous avons commencé cette expérience en analysant de le modèle SIR simple. En utilisant les mêmes paramètres (probabilité d'infection et ensembles de graines) calculés au chapitre 3 nous avons simulé ce modèle sur tous les graphes de la séquence. La structure des cascades simulées a été essentiellement la même pour tous les graphes sauf le graphe de base, qui n'a pas la même distribution de degrés que les autres graphes. Ce résultat suggère que cette propriété a eu un impact majeur pour ce modèle et pour l'échelle de temps de simulation considérée. Ensuite, nous avons effectué la même analyse avec le modèle SI temporel (c'est-à-dire, com-

portant un temps inter-contagion). Avec les paramètres calculés précédemment, au chapitre 4, nous avons observé que le modèle de SI temporel avec comportement des pairs homogène est insensible à l'ajout des propriétés topologiques complexes. En d'autres termes, les cascades simulées présentent le même profil, en dépit de l'ajout successif des propriétés topologiques complexes. En revanche, en simulant le modèle de SI temporel avec le comportement hétérogène des pairs, nous avons constaté que ce modèle est très sensible à la distribution des degrés des graines. En effet, cette propriété a été responsable de la différence de structure majeure au niveau des cascades (l'impact des autres propriétés a été mineur). Étant donné que ce modèle SI temporel a été le modèle le plus réaliste dans nos expériences, cela souligne l'importance de la distribution des degrés des graines à la structure des cascades simulées avec un temps borné.

Appendix B

Abstract

Understanding information diffusion on complex networks is a key issue from a theoretical and applied perspective. Epidemiology-inspired SIR models have been proposed to model information diffusion. Recent papers have analyzed this question from a data-driven perspective, using on-line diffusion data. We follow this approach, investigating if epidemic models, calibrated with a systematic procedure, are capable of reproducing key structural properties of spreading cascades.

We first identified a large-scale, rich dataset from which we can reconstruct the diffusion trail and the underlying network. Secondly, we examine the simple SIR model as a baseline model and conclude that it was unable to generate structurally realistic spreading cascades. We extend this result examining model extensions which take into account heterogeneities observed in the data. In contrast, similar models which take into account temporal patterns (which can be estimated with the interaction data) generate more similar cascades qualitatively. Although one key property was not reproduced in any model, this result highlights the importance of temporal patterns to model diffusion phenomena.

We have also analyzed the impact of the underlying network topology on synthetic spreading cascade structure. We have simulated spreading cascades in similar conditions as the real cascades observed in our dataset, namely, with the same time constraints and with the same "seeds". Using a sequence of uniformly random graphs derived from the real graph and with increasing structure complexity, we have examined the impact of key topological properties for the models presented previously. We show that in our setting, the distribution of the number of neighbors of seed nodes is the most impacting property among the investigated ones.

Keywords: information diffusion, spreading cascade, SIR, complex networks.