Inadequacy of SIR Model to Reproduce Key Properties of Real-world Spreading Cascades: Experiments on a Large-scale P2P System

Daniel F. Bernardes, Matthieu Latapy, Fabien Tarissan
LIP6 – CNRS and Université Pierre et Marie Curie / Paris 6
4 Place Jussieu, 75252 Paris cedex 05 – France
Email: firstname.lastname@lip6.fr

***Abstract*—Understanding the spread of information on complex networks is a key issue from a theoretical and applied perspective. Despite the effort in developing theoretical models for this phenomenon, gauging them with large-scale real-world data remains an important challenge due to the scarcity of open, extensive and detailed data. In this paper, we explain how traces of peer-to-peer file sharing may be used to this goal. We reconstruct the underlying social network of peers sharing content and perform simulations on it to assess the relevance of the standard SIR model to mimic key properties of real spreading cascades. First we examine the impact of the network topology on observed properties. Then we turn to the evaluation of two heterogeneous extensions of the SIR model. Finally we improve the social network reconstruction, introducing an affinity index between peers, and simulate a SIR model which integrates this new feature. We conclude that the simple, homogeneous model is insufficient to mimic real spreading cascades. Moreover, none of the natural extensions of the model we considered, which take into account extra topological properties, yielded satisfying results in our context. This raises an alert against the careless, widespread use of this model.**

I. Introduction

Diffusion phenomena in complex networks – such as the spread of virus on contact networks, gossip on social networks and files in peer-to-peer (P2P) networks – have spawned an increasing interest in recent years. The boost of computer networks and online social network platforms offers data and new insights on these phenomena in large scale networks; the possibility to validate and refine current models might lead to breakthroughs in the field.

Although large scale diffusion phenomena have always attracted considerable interest, it has been historically challenging to obtain open, extensive and detailed real-world data at this level. Despite this obstacle, diffusion models emerged, notably in epidemiology. The early models, both discrete and continuous (see [1], [2] for a survey), focused primarily on *macroscopic* aspects of diffusion – such as the evolution of the number of infected individuals in a population – overlooking the *microscopic* dynamic of the epidemic – i.e., how (by whom) individuals become infected. The advent of network analysis in various contexts has pushed for a more detailed description of the diffusion process. Indeed, models based on the precise interactions of individuals on a network have blossomed in sociology [3], computer science [4] and economics [5], among others. New epidemic models inspired by the classical approaches featuring a detailed dynamic description in the context of networks also appeared (see [6] for a survey). In particular the network version of the SIR model (henceforth called simply SIR model) and derivates have established themselves as reference models in the study of information diffusion [7].

In this context, we have seen theoretical developments of these models [8], [9], focusing particularly in their asymptotic behavior. A number of applications of such models were also explored [10], [11], including works investigating relevant properties of epidemic models on real networks [12]. However, as pointed out in [13], assessing the pertinence of such models to describe real-world phenomena is critical and empirical studies featuring *real spreading data* are key. Since network-based epidemic models are based on local rules of transmission which take into account the network topology, in order to validate these models one needs a comprehensive empirical spreading trace, consisting of (1) detailed chronological data of who transmitted the information to whom and (2) data describing the underlying network on which the diffusion process takes place. In large epidemic bursts the available data often provides the evolution of an aggregate quantity (such as the number of touched individuals) but rarely uncover the local trail of the epidemic at an individual level. Data mining in computer networks can help providing detailed information at a large scale [14], [15], [16], [17]. In this framework, works typically feature records of diffusion events at an individual level but lack the complete information of the underlying network on which the diffusion takes place – see discussion in [18]. The present paper analyses the relevance of the SIR model for real-world diffusions, using data obtained measuring file sharing activity on a peer-to-peer network. Our framework allows one to take advantage of this rich dataset to obtain both the real spreading data (the detailed diffusion trail) and the underlying network.

This paper extends the results in [19]. It begins with a description of our dataset and framework in section II, in

which we calculate statistics concerning peer activity and file sharing and in which we define spreading cascades. In section III we construct the underlying social network of peers from the diffusion trace. In the following two sections, we confront the SIR model and extensions to the real data. More precisely, in section IV we simulate the spreading of files as a standard SIR process and compare it with the observed spreading; we also investigate the interplay between this process and structural properties of the underlying network where the spreading takes place. In section V we examine the spreading pattern when we modify the SIR model to account for heterogeneity in the behavior of the peers and in the popularity of files. In section VI we present a novel approach which consists in simulating an SIR derived models on an enhanced reconstruction of the underlying social network. This reconstructed network is made possible using an affinity index for each couple of peers. We conclude the paper with future work perspectives.

II. Dataset and Framework

The data used in this study comes from file sharing in an eDonkey server, obtained from a measurement of eight hours of activity (akin to [20]). In this setting, peers query the eDonkey server indexing files and for each file they get a list of available peers in the network possessing the requested file. Next, peers contact potential providers directly and transmission between them ensues. Our dataset is a collection of answers to these queries, encoded as 4-tuples of integers in the following format: (t, P, C, F), where capital letters represent unique ids (e.g. in Fig. 1). Each tuple accounts for a query made at time t of the file F by the peer C, satisfied by the peer P – that is, P provided F to the peer C at time t. Let $\mathbf{D}$ be the set of all recorded tuples, $\mathcal{P}$ the set of all peers in these tuples and $\mathcal{F}$ the set of all files exchanged. In our dataset we have $|\mathcal{P}| = 1\ 908\ 500$ peers, $|\mathcal{F}| = 801\ 280$ files and $|\mathbf{D}| = 22\ 944\ 800$ file transfers.

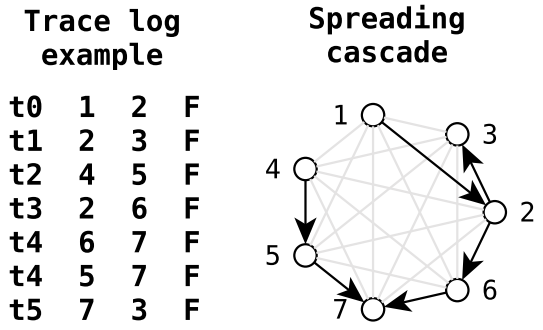


Fig. 1. Trace log example with corresponding spreading cascade in black and underlying network in light gray.

A. File sharing

The trace $\mathbf{D}$ naturally induces a relationship between files and peers (who request or provide them), which we explore descriptively in this section. We begin encoding this relationship between the disjoint sets of peers $\mathcal{P}$ and $\mathcal{F}$ files in a bipartite graph $\mathcal{B} = (\mathcal{P}, \mathcal{F}, \mathcal{A})$ as follows. Let $(t, P, X, F) \in \mathbf{D}$ be a recorded transmission of the file F by the peer P to some peer X at some time t, which we denote simply by $(\cdot, P, \cdot, F)$. Likewise, let $(\cdot, \cdot, P, F) \in \mathbf{D}$ be a recorded transmission of the file F to the peer P, provided by some peer at some time instant. Then:

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

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. The degree distribution of these sets in $\mathcal{B}$ (constructed from the P2P trace) are plotted in Fig. 2a. In order to estimate the typical number of interested peers per file we have computed the median degree of files in the bipartite graph, 5, and the average degree, 14.73, with standard deviation 34.74. Likewise, we have calculated the same statistics for peers, to estimate the number of files commonly shared by peers: its median degree in the bipartite graph is 3 and the average degree is 6.19, with corresponding standard deviation 12.66. The degree distribution of both peers and files is however heterogeneous and mostly concentrated on small values; all degree values for peers and files remain below 10^4.

Another important aspect of our P2P trace in terms of sharing is the abundance of *free-riders* – that is, peers who benefit of shared files in the system, but who do not share back. This characteristic is well known in the P2P literature and has been observed elsewhere [21]. In our dataset, while most peers are clients (i.e., have requested a least one file) only 4.33% of them have supplied files.

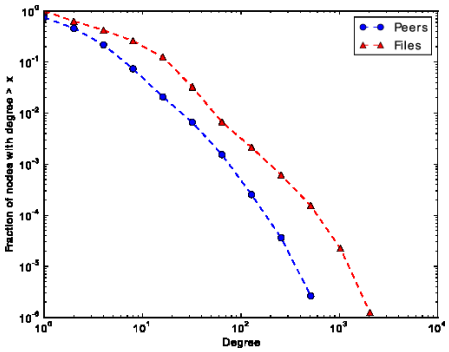
B. Spreading cascades

In this work we analyze the *spreading cascade* representing the diffusion of each file in the P2P network. 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

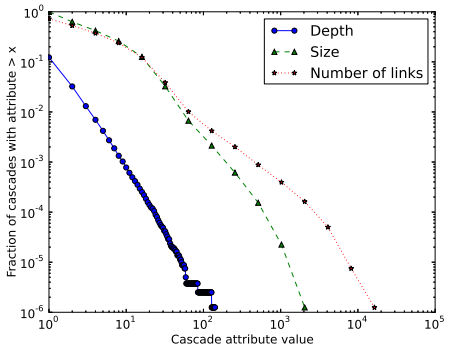
$$\mathcal{P}_F = \{P \in \mathcal{P} : (P, F) \in \mathcal{A}\}$$

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

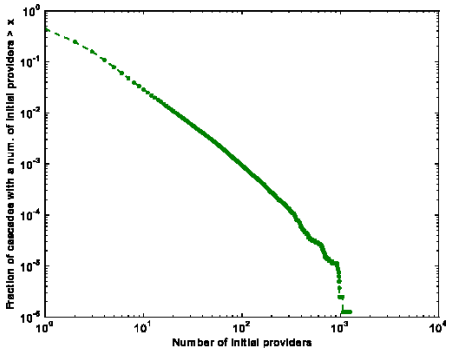
A client requesting a file may receive a response from potentially several providers simultaneously, which implies



(a) Complementary cumulative degree distributions of peers and files on the bipartite graph $\mathcal{B}$.



(b) Complementary cumulative distribution of cascade attributes (depth, size, number of links).



(c) Complementary cumulative distribution of the number of initial providers in the spreading cascades.

Fig. 2. Spreading statistics from the observed diffusion trace.

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).

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. An example of observed trace and constructed spreading cascade is given in Fig. 1: the spreading cascade has size 7, depth 3 and 6 links.

From the P2P trace log we have constructed the spreading cascades for each observed file and calculated the above mentioned features. The distribution of these cascade features is presented in Fig. 2b. 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 number of nodes observed in our dataset.

C. Initial providers

Another relevant spreading data concerns the *initial providers* 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 cascade, 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:

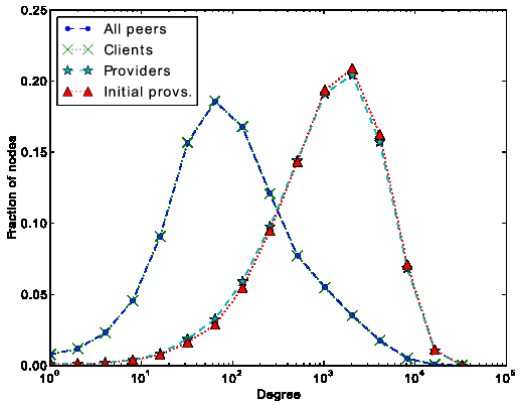
$$\mathcal{I}_F = \{P \in \mathcal{P} : (t, P, \cdot, F) \in \mathbf{D}, P \notin \mathcal{C}_F(t)\}$$

Plotting the complementary cumulative distribution of the number of initial providers for the spreading cascades (Fig. 2c) 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.

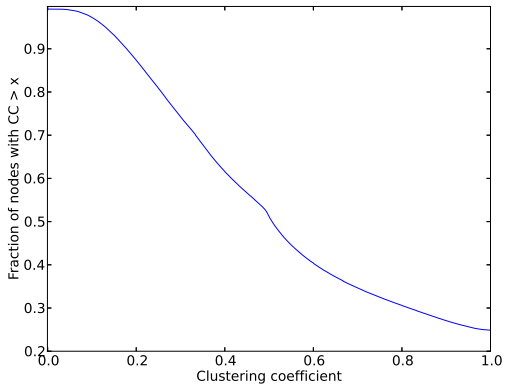
III. Social network structure

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 [18] it is challenging to reconstruct the network on which the diffusion takes place. One strategy to unfold this network is to explore relations among peers and their common shared files. Such strategy was hinted in [22] and developed more substantially in [23], [24], [19], [25]. We follow this approach to reconstruct the underlying social network as well.

Focusing on information content diffusion among peers, it is natural to consider the *interest graph* in which each



(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.

Fig. 3. Interest graph statistics

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

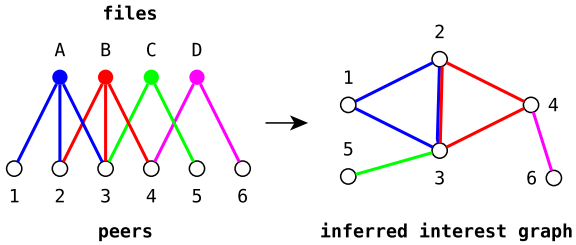


Fig. 4. Interest graph as a projection of the bipartite graph of peers and files constructed from the trace **D**.

It is beyond doubt extremely difficult in a large scale interaction network to know precisely whether any two individuals have a common interest. Nonetheless, it is possible to approximate this graph using the data in **D**: the inferred interest graph is given by the projection $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ of $\mathcal{B}$ on $\mathcal{P}$, connecting the peers who belong to the neighborhood of a common file in the bipartite graph, for each file:

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

See example in Fig. 4. For the sake of readability the inferred interest graph will be henceforth called simply *interest graph*.

The interest graph obtained from the observed bipartite graph (as explained above and in Fig. 4) has a single giant component containing essentially all nodes (99.99%) and density 2.62×10^{-4}. In Fig. 3a 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 (Fig. 3a). 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 calculated the clustering coefficient [26] of the peers in the interest graph (Fig. 3b): 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.

We close this section with a brief summary: using the introduced framework, we were able 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.

IV. Simple SIR model

As mentioned in the introduction, we have decided to investigate the file spreading in the light of the simple SIR model. In our setting, 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 denoted *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 sending the file.

Supposing the observed diffusion trace was the result of such a simple SIR epidemic we may estimate the spreading parameter p. Each neighbor-to-neighbor 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 upon a single parameter, namely the spreading probability p, we have fully characterized it with the preceding estimation.

A. The underlying network influence

The goal of simulating the standard SIR model and comparing the simulated cascades with the observed ones is primarily to assess how realistic this model would perform on the interest graph, in terms of size, depth and number of links of the spreading cascades. Note that by realistic, we mean able to reproduce the characteristics of the data. Indeed, although the data used in this study can be partial and/or biased, the present work is independent from the quality of the data itself. Indeed, the problem of improving the measurement process is different from the one of identifying relevant models able to exploit the features observed in the data, which is the focus of this paper. This means that when we further show the ability of the models to reproduce (or not) the characteristics of real traces, it has to be understood as the ability of reproducing the characteristics as *observed* in the data, with their flaws. Another approach is to apply detection techniques (such as [27]) in order to remove abnormal events from the raw data before using the modeling techniques presented in this paper. Although it could improve the quality of the data, it would at the same time obfuscate our conclusions as it adds another step which interferes in the analysis process. Thus, promising as it seems, we leave this approach for a further study.

Secondly, we wish to compare the results with simulations on random networks to understand the role of the network topological structure on the shape of the spreading cascades generated with the SIR model. With this aim, we have considered the spreading of files in a sequence of random networks derived from the interest graph, with increasing topological complexity (Fig. 5). More precisely we begin considering an Erdös-Rényi (ER) random graph with the same density 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 [8]. 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 [28]. 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.

Fig. 5. Increasingly realistic random graphs derived from the interest graph.

B. File spreading simulation

Combining the network topology, the initial condition information (the list of initial providers $\mathcal{I}_F$ calculated for each file F) and the calibrated spreading parameter $\hat{p}$ we can proceed to the simulations for each underlying network: 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 will infect each of its neighbors with probability $\hat{p}$, becoming non-interacting afterwards. The epidemic continues as long as there are active 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 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.

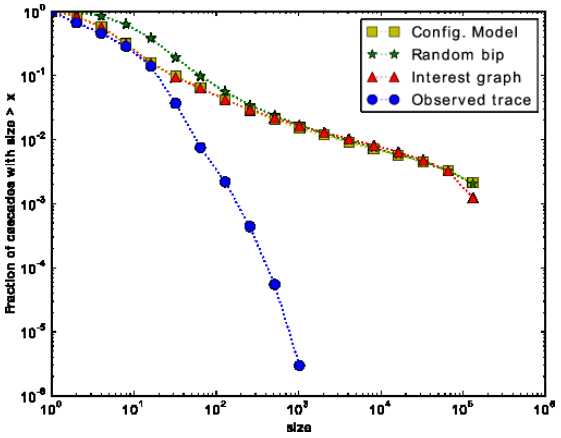
We have generated populations of simulated cascades for each underlying network and constraint (on depth and size). We have performed $801\ 280$ file spreading simulations (one for each file in $\mathcal{F}$) for each network and have selected every simulated file spreading cascade which attained the 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 Fig. 6a 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 Fig. 6c 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 (Fig. 6b) the number of links distribution resembles closely the size distribution (Fig. 6a). 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 (Fig. 6d), 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 simple, featuring depth equal to one and correspondingly small size.

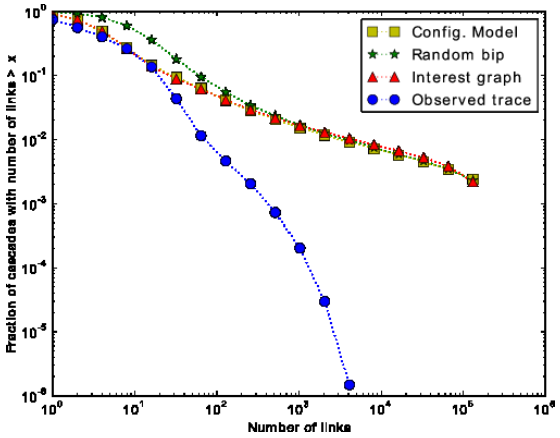
To sum up, we have compared simple topological properties of real spreading cascades and simulated cascades from a calibrated SIR model, with comparable depth and size. We have observed that simulated cascades are relatively "wider" whereas real cascades are relatively "elongated", that is, real cascades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

V. Heterogeneous SIR models

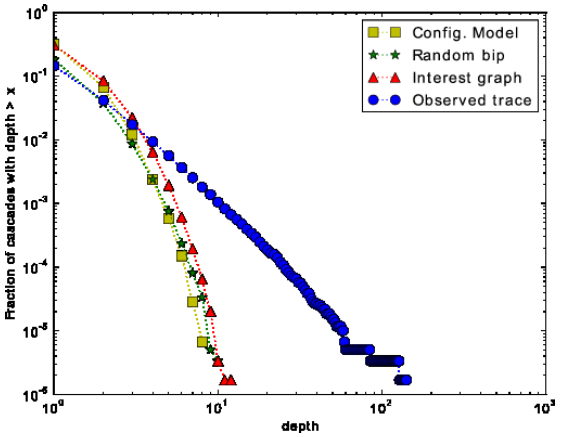
In the previous section we have examined the adequacy of the simple SIR model to generate verisimilar file spreading cascades. We have also inspected the interplay between the underlying network and the model simulating file spreading in different networks. Given the generality and simplicity of the 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 model 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 a single underlying network (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.



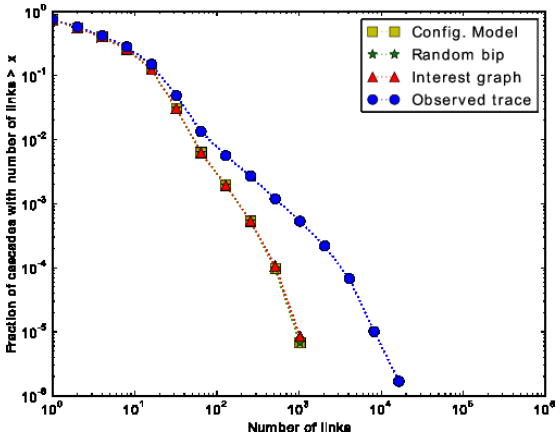
(a) Size of cascades with fixed depth. Curves corresponding to the interest graph and CM superposed.



(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.

Fig. 6. Simulation of file spreading on different underlying networks: complementary cumulative distribution of cascade properties

A. 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 Fig. 7a 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.

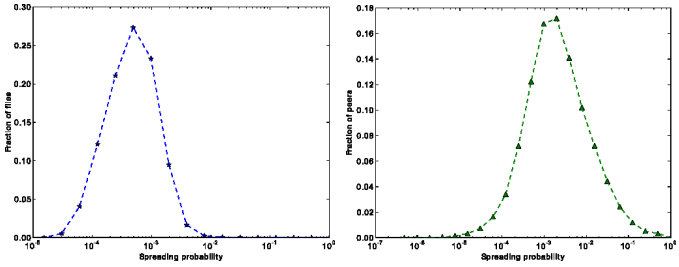
B. Peer behavior

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 (Fig. 7b). 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 section II. 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 calculated for the homogeneous SIR.



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

Fig. 7. Heterogeneous spreading parameter distributions

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 Fig. 8: we have plotted the complementary cumulative distributions of the spreading cascade depth, size and number of links. Imposing a constrain on the depth for the simulated cascades and comparing their size (Fig. 8a) 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 latter. What is remarkable, however, is the agreement among all the simulated cascade distributions – curves superposed in Fig. 8a. Next, if we fix the size for the simulated cascades and examine their depth (Fig. 8c), we are faced with 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 Fig. 8b and Fig. 8d reveals the pattern observed previously, namely that the observed spreading cascades are typically denser than corresponding simulated cascades.

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 SIR models compared, notwithstanding the particularities of each model.

VI. Weighted interest graph

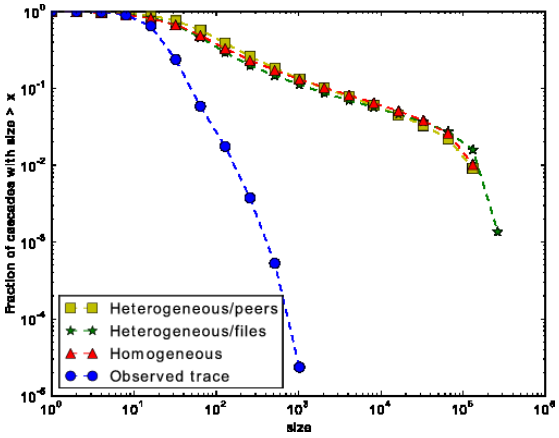
In the previous section we have examined SIR model extensions that take into account heterogeneous aspects of peers and files with the goal of generating more realistic spreading cascades. Another approach is to keep the simple SIR model and enrich the social network inference. In this section we will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

A. File spreading simulation

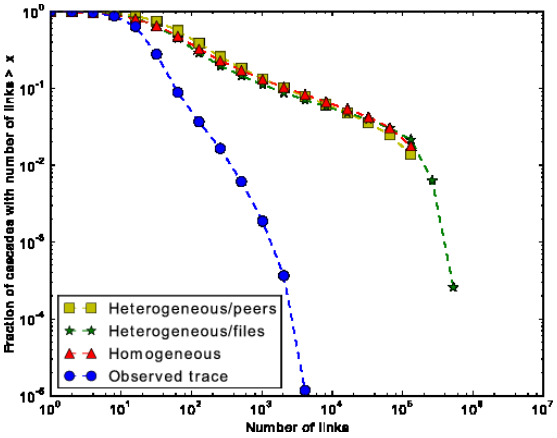
In concrete terms, our affinity score between two peers will be defined by the number of common files peers shared or provided. Indeed, instead of approximating the interest graph by the simple projection of $\mathcal{B}$ on $\mathcal{P}$, we consider a richer inferred interest graph $\mathcal{G} = (\mathcal{P},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

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

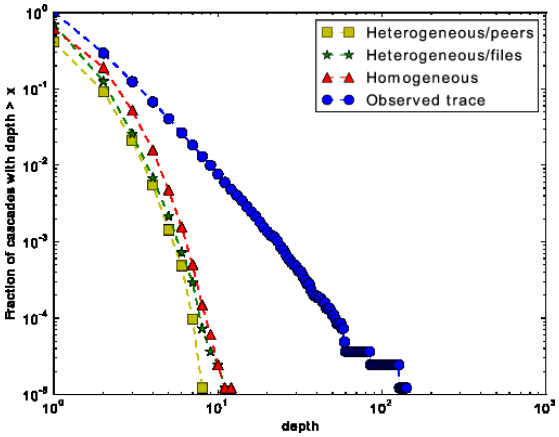
$$\mathcal{W}(P, P') = |\{F \in \mathcal{F} : (P, F) \in \mathcal{A} \wedge (P', F) \in \mathcal{A}\}|$$



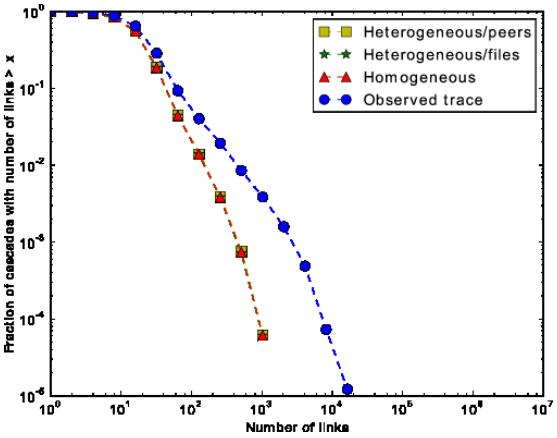
(a) Size of cascades with fixed depth. Curves corresponding to the simulations are superposed.



(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.

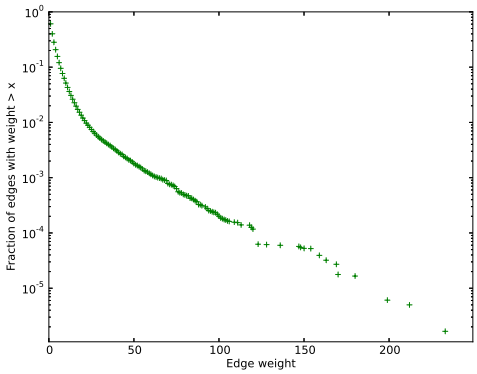
Fig. 8. Simulation of file spreading on the interest graph with different SIR processes: complementary cumulative distribution of cascade properties

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. In Fig. 9a we have plotted the distribution of weight values in the interest graph: it is heterogeneous, with weights ranging from 1 to 303 and such that the vast majority of edges feature 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.

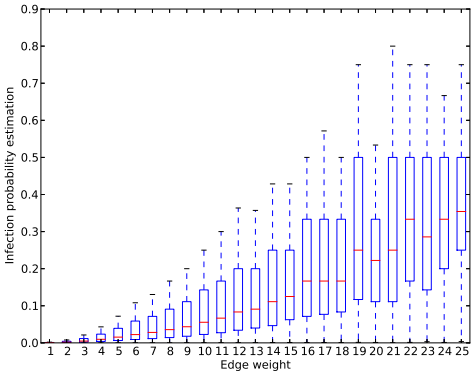
B. Diffusion models

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*, *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 suppose that a peer P will be 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



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



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

Fig. 9. 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.

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 IV, V. 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 Fig. 9 (right) we have plotted the estimator values for all weights from 1 to 25 in terms of box plots.

Following the approach in [29], 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. Fig. 9a. 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.

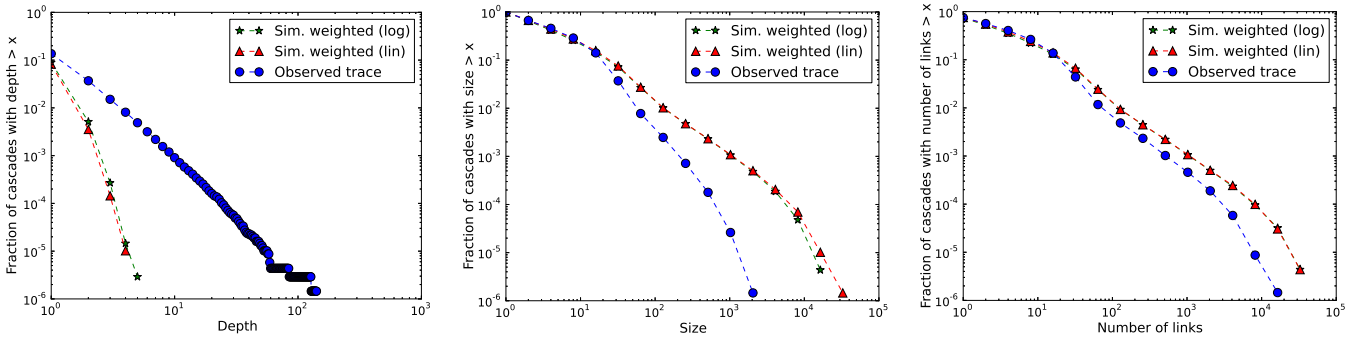
C. File spreading simulation

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 Fig. 10 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 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 [30], in the sense that they are relatively elongated compared to simulated cascades obtained with these contagion models.

VII. Conclusion and perspectives

We have presented a large-scale dataset from a real-world peer-to-peer network, featuring diffusion of files among peers. We have proposed a framework to study this dataset which allows us to obtain, simultaneously, the interest graph of peers – where the diffusion of content takes place – and the spreading cascade. Guided by simulations we have examined spreading cascades generated by the simple SIR model and have analyzed the interplay between this model and the network topology. We concluded that simulated file diffusions do not capture key qualitative properties of the observed spreading cascades. Furthermore, in terms of the studied



(a) (b) (c)

Fig. 10. 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.

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 the addition of a clustering coefficient constraint on the random graph did not change the properties of the spreading cascade qualitatively.

The SIR model is an attractive choice to model the information spreading in complex networks: it was inspired by classical epidemiological models, it is based upon few assumptions and it can be characterized with few parameters. This flexibility and simplicity explain its popularity as a contagion model, but these characteristics are also its weakness when used in specific contexts. In this sense, the results of section IV, mentioned above, are not entirely surprising. What is surprising, though, is that 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.

In sum, these results suggest that this model is not suited to describe information spreading in our context. That is, not even the natural extensions of this model, related to key observed features of real spreading cascades, offer a better alternative in terms of the properties we have investigated. It is evidently hard to demonstrate that there is no possible modification of this model capable of describing file spreading cascade profiles in P2P systems, but our results show that this model is unlikely to describe spreading cascades generally, as it is commonly taken for granted. In this sense, our work raises a cautionary message against the careless, widespread use of this model.

Although the spreading cascade modelling seems to be more context-dependent than currently thought, the precise role of context in the choice of model and its parameters remains open. In our case, we have focused primarily on the interplay between the diffusion process and the network structure and have neglected other potentially important aspects in this context. Two aspects in particular could explain why SIR-based models fail to reproduce the profile of spreading cascades.

The first one concerns the time, which is not directly addressed in such models except from a logical point of view. This aspect is however strongly related to the order in which the underlying graph is explored during a contagion. The logical nature of the time adopted here is similar to a breadth-first-search exploration which yields short depth structures. This explains particularly the inadequacy observed when the depth is involved in the evaluation of the model, such as in Fig. 6a, Fig. 6c, Fig. 8a, Fig. 8c, and Fig. 10a. In contrast, figures corresponding to size and number of links show a clear improvement of the model efficiency. Thus, it seems very promising to exploit more deeply the temporal information in our dataset. One possible way would be to take into account the dynamic aspect of the social network by filtering the interest graph with pairs of peers that have been present at the same time in the network. Another way would be to incorporate time-related behavior to the contagion dynamics, that is to add a new features in the model itself that account for such temporal patterns. Though promising, we leave such approaches for further studies.

The second aspect, which is by far more fundamental as it questions the nature of the model itself, relies on the fact that epidemic models are based on *"push"* dynamics whereas peers in P2P systems tend to *"pull"* content from each others. This might call for a fundamental perspective change on the dynamics of the process. In particular, adoption/threshold models [13], [3] could be more pertinent in this case: we also plan to evaluate this possibility in the future.

ACKNOWLEDGMENT

This work is partly funded by the European Commission through the FP7-FIRE project EULER (Grant No.258307) and by the City of Paris *Émergence* program through the DiRe project.

REFERENCES

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

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

[3] M. Granovetter, "Threshold Models of Collective Behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

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

[5] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton University Press, 2008.

[6] A. Barrat, M. Barthlemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. New York, NY, USA: Cambridge University Press, 2008.

[7] M. Draief and L. Massoulié, *Epidemics and rumours in complex networks*, ser. London Mathematical Society lecture note series. Cambridge University Press, 2010, no. 369.

[8] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

[9] B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, and C. Faloutsos, "Got the Flu (or Mumps)? Check the Eigenvalue!" Apr. 2010.

[10] K. Hosanagar, P. Han, and Y. Tan, "Diffusion models for peer-to-peer (p2p) media distribution: On the impact of decentralized, constrained supply," *Info. Sys. Research*, vol. 21, no. 2, pp. 271–287, Jun. 2010.

[11] K. Leibnitz, T. Hossfeld, N. Wakamiya, and M. Murata, "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*, ser. BioADIT'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 322–329.

[12] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, "The role of the airline transportation network in the prediction and predictability of global epidemics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 7, pp. 2015–2020, 2006.

[13] J.-P. Cointet and C. Roth, "How realistic should knowledge diffusion models be?" *Journal of Artificial Societies and Social Simulation*, vol. 10, no. 3, p. 5, 2007.

[14] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading Behavior in Large Blog Graphs," Apr. 2007.

[15] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose, "Implicit structure and the dynamics of blogspace," in *World Wide Web Conference Series*, 2004.

[16] J. L. Iribarren and E. Moro, "Impact of Human Activity Patterns on the Dynamics of Information Diffusion," *Physical Review Letters*, vol. 103, no. 3, pp. 038 702–+, Jul. 2009.

[17] M. Cha, J. Pérez, and H. Haddadi, "The spread of media content through blogs," *Social Network Analysis and Mining*, vol. 2, no. 3, pp. 249–264, 2012.

[18] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Trans. Knowl. Discov. Data*, vol. 5, no. 4, pp. 21:1–21:37, Feb. 2012.

[19] D. F. Bernardes, M. Latapy, and F. Tarissan, "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, 2012, istanbul, Turkey; 2012-08-26 – 2012-08-29.

[20] F. Aidouni, M. Latapy, and C. Magnien, "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*, 2009, pp. 1–5.

[21] E. Adar and B. Huberman, "Free riding on gnutella," *First Monday*, vol. 5, no. 10-2, 2000.

[22] S. B. Handurukande, A.-M. Kermarrec, F. Le Fessant, L. Massoulié, and S. Patarin, "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*, ser. EuroSys '06. New York, NY, USA: ACM, 2006, pp. 359–371.

[23] M. Latapy, C. Magnien, and N. D. Vecchio, "Basic notions for the analysis of large two-mode networks," *Social Networks*, vol. 30, no. 1, pp. 31 – 48, 2008.

[24] A. Iamnitchi, M. Ripeanu, E. Santos-Neto, and I. Foster, "The small world of file sharing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 7, pp. 1120–1134, Jul. 2011.

[25] O. Allali, L. Tabourier, C. Magnien, and M. Latapy, "Internal links and pairs as a new tool for the analysis of bipartite complex networks," *Social Network Analysis and Mining*, vol. 3, no. 1, pp. 85–91, 2013.

[26] A. Barrat, M. Barthlemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. New York, NY, USA: Cambridge U. Press, 2008.

[27] H. Sencan, Z. Chen, W. Hendrix, T. Pansombut, F. H. M. Semazzi, A. N. Choudhary, V. Kumar, A. V. Melechko, and N. F. Samatova, "Classification of emerging extreme event tracks in multivariate spatio-temporal physical systems using dynamic network structures: Application to hurricane track prediction," in *IJCAI*, 2011, pp. 1478–1484.

[28] J.-L. Guillaume and M. Latapy, "Bipartite structure of all complex networks," *Inf. Process. Lett.*, vol. 90, no. 5, pp. 215–221, 2004.

[29] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[30] D. Liben-Nowell and J. Kleinberg, "Tracing information flow on a global scale using internet chain-letter data," *Proceedings of the National Academy of Sciences*, vol. 105, no. 12, pp. 4633–4638, 2008.