Relevance of SIR Model for Real-world Spreading Phenomena: Experiments on a Large-scale P2P System

Daniel Bernardes, Matthieu Latapy, Fabien Tarissan

Relevance of SIR Model for Real-world Spreading Phenomena: 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 also perform simulations to assess the relevance of the standard SIR model to mimic key properties of spreading cascade. We examine the impact of the network topology on observed properties and finally turn to the evaluation of two heterogeneous versions of the SIR model. We conclude that all the models tested failed to reproduce key properties of such cascades: typically real spreading cascades are relatively "elongated" compared to simulated ones. We have also observed some interesting similarities common to all SIR models tested.**

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 detailed interactions of agents 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], [7] for a survey). In particular the network version of the SIR family of models has established itself as reference model in the study of information diffusion [8], [9], [10], [11], [12].

In this context, assessing the pertinence of such models to describe real-world data is critical. In order to validate this model 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. Indeed, the network version of the SIR model (henceforth called simply SIR model) is based on local rules of transmission which take into account the network topology. In large epidemic bursts the available data often provides the evolution of large aggregate quantity, such as the number of touched individuals, but rarely uncover the local trail of the epidemic. Conversely, other empirical studies feature transmission events, but lack complete information of the underlying network structure on which the diffusion takes place [13], [14]. This work analyses the relevance of the SIR model for real-world diffusions, using data obtained measuring the activity on a peer-to-peer file sharing network. This rich dataset allows one to reconstruct both the underlying network and the detailed diffusion trail at a remarkable scale.

We begin with a description of our dataset and framework in section II. In section III we define the spreading cascade. Next, in section IV, we simulate the spreading of files as a standard SIR process and confront 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. We conclude the paper with of 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 six hours of activity (akin to [15]). In this setting, peers query the eDonkey

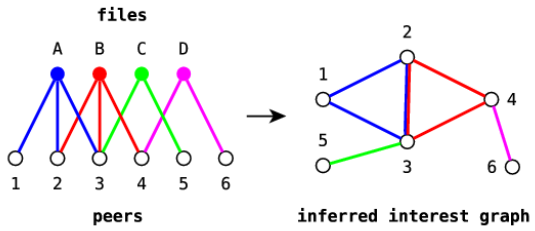


Fig. 1. Interest graph as a projection of the bipartite graph of peers and files contructed from the trace $\mathbf{D}$.

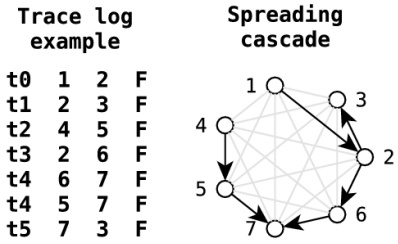


Fig. 2. Trace log example with corresponding spreading cascade in black and underlying network in light gray.

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. This 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. 2). 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 has 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.

A. Underlying network

The trace $\mathbf{D}$ naturally induces a relationship between files and peers (who request or provide them), which we encode in a bipartite graph $\mathcal{B} = (\mathcal{P}, \mathcal{F}, \mathcal{A})$ on the disjoint sets of peers $\mathcal{P}$ and $\mathcal{F}$ files respectively. 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. Hence:

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

To study the diffusion, it is necessary to define the underlying network on which spreading takes place. Focusing on information 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 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.

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 $\mathbf{D}$: 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:

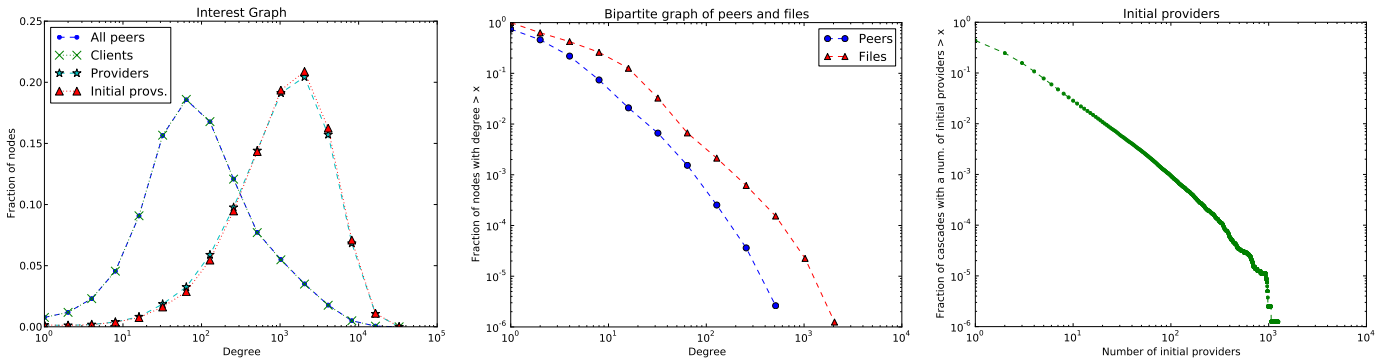
$$\mathcal{E} = \{(P, P') \in \mathcal{P} \times \mathcal{P} : \exists F \in \mathcal{F}, (P, F) \in \mathcal{A} \wedge (P', F) \in \mathcal{A}\}$$

See example in Fig. 1. For the sake of readability the inferred interest graph will be henceforth called simply *interest graph*.

B. Observed network structure

We begin examining properties of the bipartite graph $\mathcal{B}$ constructed from the P2P diffusion trace. In order to estimate the typical number of interested peers per file we have calculated the median degree of the files in the bipartite graph, 5, and the average degree, 14.73, with standard deviation of 34.74. Likewise, we have calculated the same statistics for the 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 of 12.66. The degree distribution of both peers and files is however heterogeneous (Fig. 3b) and mostly concentrated on small values; all degree values for peers and files remain below 10^4.

The interest graph obtained from the observed bipartite graph (as explained above and in Fig. 1) has a single giant component containing almost 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



(a) Degree distributions on the interest graph. Superposed curves: all peers and clients, providers and initial providers

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

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

Fig. 3. Properties of the underlying network and observed spreading cascades

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. They account for 4.33% of the peers in $\mathcal{P}$. 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 a particular subset of providers called *initial providers*, which will be detailed in the next section.

We close this section with a brief summary of our dataset. Using the framework introduced, we were able to reconstruct the interest graph of the peers, where 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: this revels a high proportion of *free-riders* (peers obtain files and do not share back) in the P2P-network observed. Furthermore, there is a sharp distinction between clients and providers in terms of their degree distribution.

III. Spreading in our data

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 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 undergo 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. 2: the spreading cascade has size 7, depth 3 and 6 links.

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. 3c) 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.

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. Infected nodes, in turn, spread the file to each of its neighbors, independently, with probability p and become 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

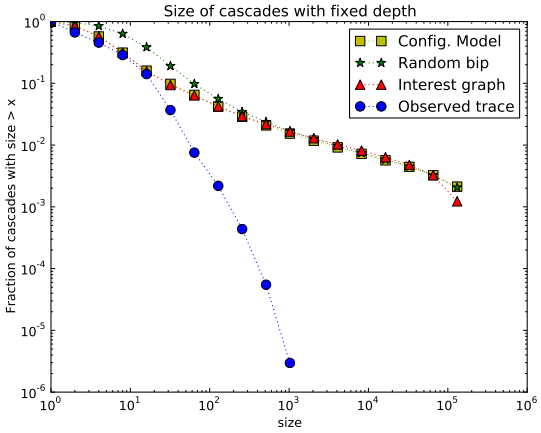
The goal of simulating the standard SIR model and comparing the simulated cascades with the observed ones is primarily to assess how realistic this model would perform on the interest graph, in terms of size, depth and number of links of the spreading cascades. Secondly, we wish to compare the results 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. More precisely we begin considering an Erdös-Rényi (ER) random graph with the same density of 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 [9]. Next we have generated a Random Bipartite (RB) graph, with the same density and degree distribution as our original bipartite graph $\mathcal{B}$ of peers and files [16]. Compared to the interest graph, the projection of this random bipartite graph 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.

B. File spreading simulation

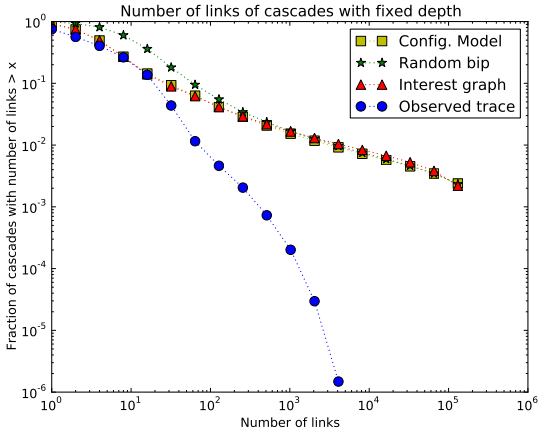
Combining the network topology, the initial condition information (the list of initial providers $\mathcal{I}_F$ calculated for each file F) and the calibrated spreading parameter $\hat{p}$ we can proceed to the simulations for each underlying network: for each F, we begin with the initial providers 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 six 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.

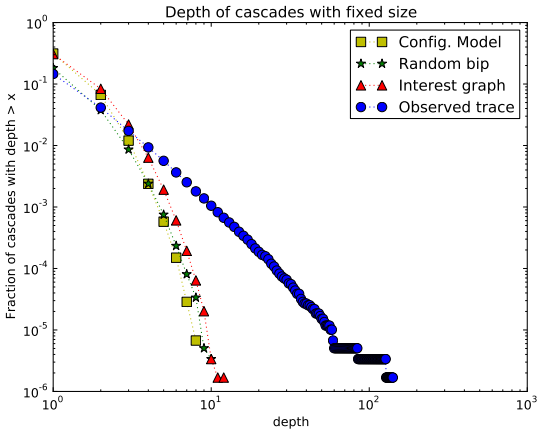
We have generated populations of simulated cascades for each underlying network and constraint (on depth and size). We have performed $801\,280$ file spreading simulations (one for each file in $\mathcal{F}$) for each network and have selected every simulated file spreading cascade which attained the depth (resp. size) of the real spreading cascade for the same file –



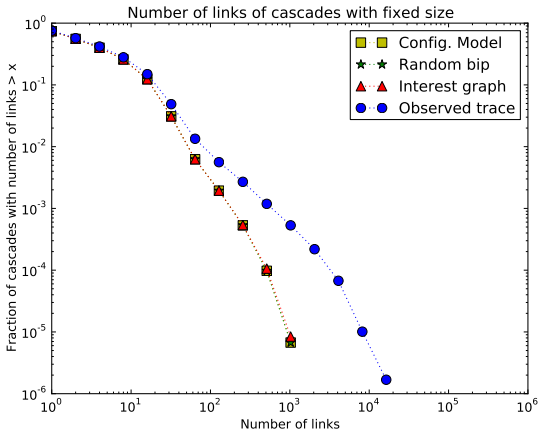
(a) Size of cascades with fixed depth. Curves corresponding to the interest graph and CM superposed.



(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. 4. Simulation of file spreading on different underlying networks: complementary cumulative distribution of cascade properties

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 of files with comparable spreading cascade depth (resp. size) on all networks but ER.

In Fig. 4a 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. 4c 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. 4b) the number of links distribution resembles closely the size distribution (Fig. 4a). 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 itselves are bigger. If, in constrast, we fix the cascade size to fit the observed cascades size (Fig. 4d), 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.

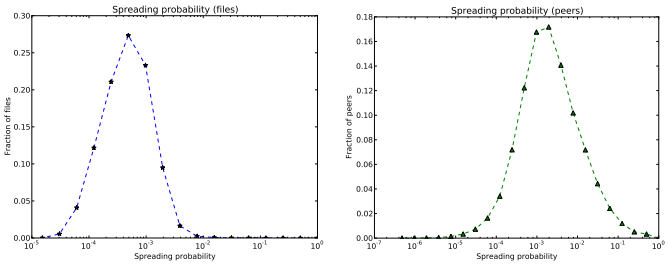
To sum up, we have compared simple topological properties of real spreading cascades and simulated cascades from a calibrated SIR model, with comparable depth and size. We have observed that simulated cascades are relatively "wider" whereas real cascades are relatively "elongated", that is, real cascades 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. In this section we perform a complementary analysis, focusing on a single underlying network and examining different extensions of the SIR model considered previously. In particular we consider two heterogeneous versions of the SIR model, characterized by a distribution of spreading probabilities, instead of a single homogeneous parameter. The natural choice in this case for the underlying network is the interest graph, which is the most complete and realistic graph among the ones tested in the previous section.

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 an artist – can change the supply and demand of a given file and consequently alter its spreading probability. The knowledge of the actual reasons that explain the heterogeneity in file popularity are irrelevant to the characterization of this model, if we know the spreading probabilities for each file, i.e., $\{p(F) : F \in \mathcal{F}\}$. 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.



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

Fig. 5. Heterogeneous spreading parameter distributions

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:

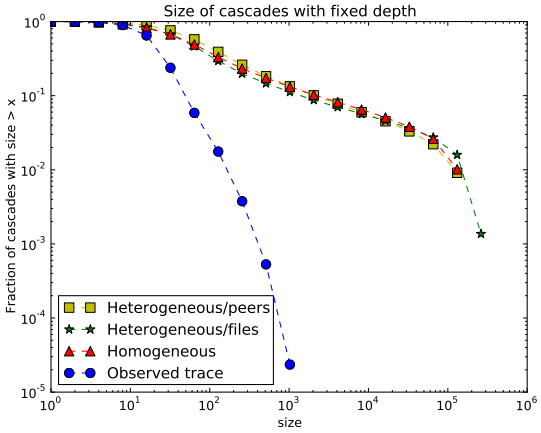
$$\hat{p}(F) = |\{(\cdot,\cdot,\cdot,F) \in \mathbf{D}\}| \;/\; \sum_{P \in \mathcal{P}_F} d(P)$$

In Fig. 5a 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.

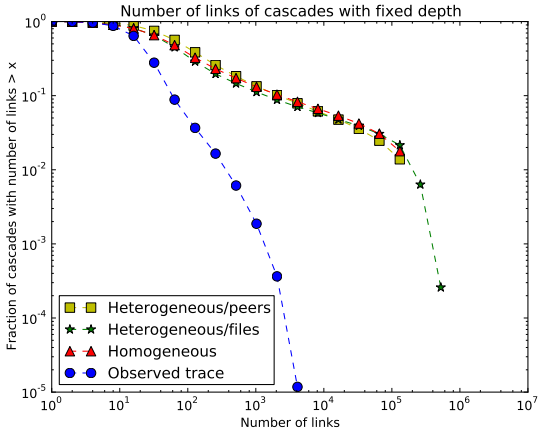
B. Peer behavior

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:

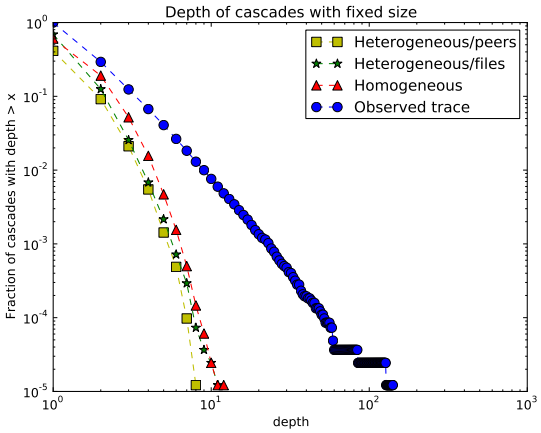
$$\hat{p}(P) = \frac{|\{(\cdot,P,\cdot,\cdot) \in \mathbf{D}\}|}{|\mathcal{F}_P| \times d(P)}$$



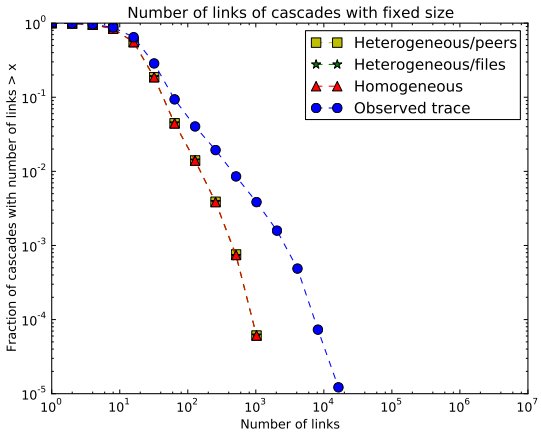
(a) Size of cascades with fixed depth. Curves corresponding to the simulations are superposed.



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

We have plotted the distribution of the positive spreading probabilities estimates in this case (Fig. 5b). 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 – 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.

C. File spreading simulation

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 the simulated cascade of the simple SIR model and the real observed cascades. In this sense, we apply the same methodology of the 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. 6: 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. 6a) 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. 6a. Next, if we fix the size for the simulated cascades and examine their depth, 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. 6b and Fig. 6d reveals the pattern observed previously, namely that the observed spreading cascades are typically denser than corresponding simulated cascades.

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

VI. 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 spreadings do not capture key qualitative properties of the observed spreading cascades. 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. Next we have focused on the spreading of files on the interest graph and studied extended versions of the SIR model featuring an heterogeneous spreading parameter. Surprisingly enough, simulated cascades using both extensions of the SIR model show similar properties as the simple homogeneous SIR model – and thus, fail to reproduce qualitative features of the observed cascades.

The SIR model is an attractive choice to model the information spreading in complex networks: it is based on classical epidemiological models, it is based upon few assumptions and can be characterized with one parameter. However, the results on this paper suggest that this model might not be suited to describe file spreading in our data. Furthermore, extensions of this epidemic model to make it more realistic, featuring heterogeneous spreading probabilities do not offer a better alternative in terms of the properties we observed. At this point, we consider two main exploration tracks. The first possibility consists in constructing a weighted interest graph, which takes into account the number of interactions (file exchanges) between peers. In this case the same analysis may be performed and a comparison with the results of this paper would be pertinent. The second possibility is to contrast epidemiological models to adoption/threshold models [8], [3].

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] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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