Examining Key Properties of Diffusion Models for Large-Scale Real-World Networks

Daniel F. Bernardes, Matthieu Latapy, Fabien Tarissan [†]

LIP6 – CNRS/UPMC, 4 Place Jussieu, F-75252 Paris cedex 05, France. Email: firstname.lastname@lip6.fr.

Understanding the spread of information on complex networks is a key issue from a theoretical and applied perspective. Despite the large effort in developing models for this phenomenon, gauging them with large-scale real-world data remains an important obstacle in the field. In this work we assess the relevance of the classic SIR model to capture key properties of spreading phenomena in real communication networks. We use a real file spreading trace in a P2P network to calibrate the model and to simulate similar diffusions. Comparing spreading cascades of real and simulated traces we observe sharp topological differences and conclude that this model fail to mimic key properties of such cascades.

Keywords: P2P networks, information diffusion, network topology

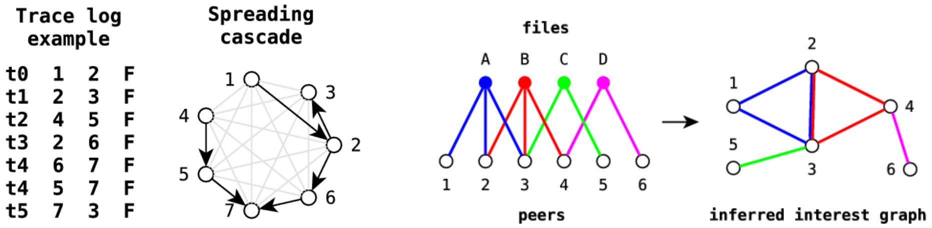
1 Introduction

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 due to the boost of computer networks and online social network platforms. Although large scale diffusion phenomena have always known considerable interest, it has been consistently challenging to obtain open, extensive and detailed real-world data at this level [IM09]. In contrast, this work is based upon a rich dataset, allowing us to reconstruct the underlying network *and* the diffusion trail at an exceptional scale. The goal of this work is twofold: to introduce a framework for studying the spread of files among users in a P2P network and to explore the interplay between this diffusion process and the underlying network. The latter is accomplished by characterizing the observed diffusion trace and by comparing this trace to model simulations. In the context of spreading in networks, cascading models are ubiquitous – particularly the SIR model, which is a standard cascading model and an archetype of several epidemics models [BBV08]. We chose to analyze file diffusions using this model, i.e., regarding the spreading of each file as an epidemic.

2 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 [ALM09]). In this setting, peers query the eDonkey server indexing files and for each requested file they get a list of available peers in the network possessing it. Next, the interested peer contacts the potential providers directly and the transmission between them ensues. This dataset is a collection of these satisfied queries, encoded as 4-tuples of integers in the following format: (t,P,C,F), where the capital letters represent unique ids (e.g. in Figure 1). Each tuple accounts for a request made at time t of the file F by the peer C, satisfied by the peer P – that is, the peer P has provided the file F to the client C at time t. Let $\mathbf{D}$ be the set of tuples constituting the spreading trace log, $\mathcal{P}$ the set of all peers in these tuples and $\mathcal{F}$ the set of all files exchanged. In this dataset we have registered $|\mathcal{P}| = 1\ 908\ 500$ peers, $|\mathcal{F}| = 801\ 280$ files and $|\mathbf{D}| = 22\ 944\ 800$ file transfers.

[†] Supported by the City of Paris through the *Emergence* project DiRe and by the European Comission through the EULER project.



```
t0  1  2  F
t1  2  3  F
t2  4  5  F
t3  2  6  F
t4  6  7  F
t4  5  7  F
t5  7  3  F
```

Figure 1: Trace log example with corresponding spreading cascade in black (left) and the interest graph as the projection of bipartite graph (right).

In order to analyze the dataset formally, we need to choose an appropriate framework. The spreading trace $\mathbf{D}$ naturally reveals a relationship between files and peers (who request or provide them), which we encode in a bipartite graph $\mathcal{B} = (\mathcal{P}, \mathcal{F}, \mathcal{A})$ with $\mathcal{A} = \{(P,F) \in \mathcal{P} \times \mathcal{F} : (\cdot, P, \cdot, F) \in \mathbf{D} \vee (\cdot, \cdot, P, F) \in \mathbf{D}\}$. To study the diffusion, it is necessary to define the underlying graph on which the spreading takes place. Since the diffusion consists of file sharing among peers, it is natural to consider the *interest graph* in which the nodes are the peers and the edges stand for a common interest of the peers. This concept is key, since 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 evidently difficult in a large scale interaction network to know precisely whether any two individuals have a common interest. Nonetheless, it is possible to infer this graph using the data in $\mathbf{D}$: the interest graph is given by the projection 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. That is, let $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ be this projection, so that $\mathcal{E} = \{(P,P') \in \mathcal{P} \times \mathcal{P} : (P,F) \in \mathcal{A} \wedge (P',F) \in \mathcal{A},\ F \in \mathcal{F}\}$. See example in Figure 1 (right).

3 Spreading in our data

In this work we investigate 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

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

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 peer received F prevents the appearance of cycles. Hence the cascade is actually a directed acyclic graph (DAG). An example of observed trace and constructed spreading cascade is given in Figure 1. Given the complexity of these cascades, we have focused our analysis on the following three key properties: size $|\mathcal{P}_F|$, number of links $|\mathcal{L}_F|$ and depth (the length of the longest path on the DAG).

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 SIR epidemic, we proceed to the estimation of 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. Given that each trial is independent and the parameter p is homogeneous for each P and F, we may estimate it by the empirical proportion of aggregated 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 the each of 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 homogeneous parameter, namely the spreading probability p, we have fully characterized it with the preceding estimation.

4 Spreading simulations

Our goal is to assess if diffusion simulations using this model to generate realistic spreading cascades, in terms of the properties described previously. Another key parameter needed to perform the simulations is the initial condition of the system. More precisely, we need to obtain the set of *initial providers* for each file F, namely the peers that possessed F prior to any transfer activity on the observed trace. 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: $I_F = \{P \in \mathcal{P} : (t, P, \cdot, F) \in \mathbf{D}, P \notin \mathcal{C}_F(t)\}$. Combining this information with the calibrated spreading parameter $\hat{p}$ we can proceed to the simulations: 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 inactive 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 relative to the cascades 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. 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.

The results are shown in Figure 2: for each cascade property (depth, size and number of links) we have plotted the complementary cumulative distribution for real cascades, depth-bounded simulations and size-bounded simulations. In the left plot, we observe a greater occurrence of small cascades in terms of depth (longest path in the spreading cascade) in the simulations in comparison with the real observed trace. The curve corresponding to depth-bounded simulations remains close to the curve corresponding to the real trace for small values of depth (cascades with depth smaller than 10); beyond this value we see a sharp cutoff – that is, few occurrences of simulated cascades of cascades with size greater than 10 and none with depth greater than 13. This breakdown in cascade depth is even more striking in size-bounded simulations, where we encounter a yet greater occurrence of small depth cascades. In the center plot we have a sharp contrast

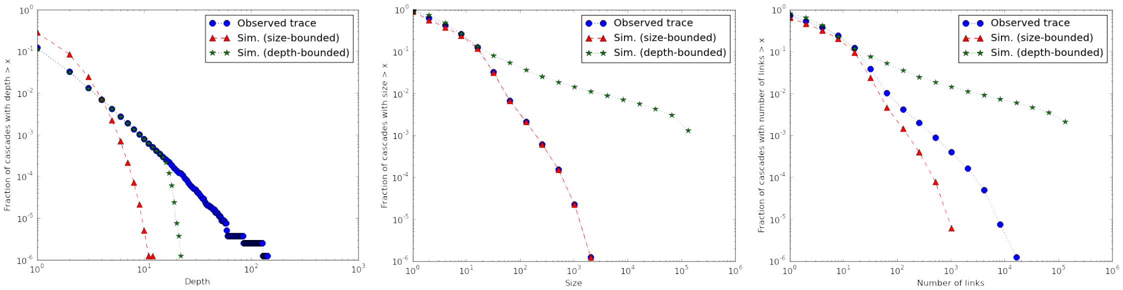


Figure 2: Spreading cascade properties: complementary cumulative distribution of depth (left), size (center) and number of links (right). Observed trace and depth-bounded simulations superposed in the center plot.

between the large size (number of nodes) of depth-bounded simulations compared to the other two curves (superposed): these point out that the SIR model features an acute growth rate in cascade size, compared to the real trace data. Indeed, the matching of the observed trace and the size-bounded simulations indicate that the cascades of the size-bounded simulation consistently attain their bound. The depth-bounded curve reinforces this diagnose, featuring cascades with a significantly bigger size, for a fixed depth. The plot on the right featuring the number of links distributions presents another interesting perspective on the spreading cascades: comparing two populations of cascades with same size (observed trace and size-bounded simulations) we conclude that real cascades are typically denser than simulated ones. Finally we note that the number of links for the depth-bounded simulation are sharply bigger than the others, which is not surprising, given that cascades in this case are also much bigger in 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.

5 Discussion

As far as the basic spreading cascade properties we have investigated are concerned, the results point to sharp divergences between the properties obtained from the real diffusion trace and the simulations. This suggests that the simple SIR model on the underlying graph, with same initial conditions and calibrated parameter $\hat{p}$ is *insufficient* to capture the basic properties of large-scale real-world diffusion processes, in regard to the basic spreading cascade properties. This calls for a significant improvement of the model, possibly taking into account more subtle aspects of the underlying graph topology than simply the node degree. One option would be to define different classes of peers and/or files each with a heterogeneous diffusion spreading parameter. Another hint would be to investigate the community structure on the graph, and have several diffusion regimes on different communities.

References

[ALM09] Frederic Aidouni, Matthieu Latapy, and Clémence 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*, pages 1–5, 2009.

[BBV08] Alain Barrat, Marc Barthlemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, New York, NY, USA, 2008.

[IM09] J. L. Iribarren and E. Moro. Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Physical Review Letters*, 103(3):038702–+, July 2009.