File Diffusion in a Dynamic Peer-to-peer Network[*]

Alice Albano
LIP6 / UPMC
4 Place Jussieu
75005 PARIS
alice.albano@lip6.fr

Jean-Loup Guillaume
LIP6 / UPMC
4 Place Jussieu
75005 PARIS
jean-loup.guillaume@lip6.fr

Bénédicte Le Grand
LIP6 / UPMC
4 Place Jussieu
75005 PARIS
benedicte.le-grand@lip6.fr

ABSTRACT

Many studies have been made on diffusion in the field of epidemiology, and in the last few years, the development of social networking has induced new types of diffusion. In this paper, we focus on file diffusion on a peer-to-peer dynamic network using eDonkey protocol. On this network, we observe a linear behavior of the actual file diffusion. This result is interesting, because most diffusion models exhibit exponential behaviors. In this paper, we propose a new model of diffusion, based on the SI (Susceptible / Infected) model, which produces results close to the linear behavior of the observed diffusion. We then justify the linearity of this model, and we study its behavior in more details.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement

Keywords

peer-to-peer, file diffusion, eDonkey protocol

1. CONTEXT

Observing and trying to predict the way a virus spreads in a population represents a major research challenge. The simplest epidemiological model is the SI (Susceptible / Infected) model, described in [14]. In this model, nodes can only take two states: susceptible or infected, and at each time step, an infected node can infect its neighbors with a fixed probability p. Many studies have already focused on diseases spreading, like [4] or [13]. Later, diffusion has also been studied in other fields such as sociology [10], economics [2] or computer science [12].

Moreover, the development of social networks in the last few years has induced the study of new diffusion types: information [10] [11] and innovation spreading [3] [5]. In this paper, we study file diffusion on a dynamic peer-to-peer network using the eDonkey protocol. Some works have recently studied file diffusion on peer-to-peer networks. In [7], the authors study the number of peers providing or looking for most popular files and they observe a linear growth for both. However, most studies are based on simulations or models: in [8], the authors propose a model in order to maximize product diffusion speed, and consider exponential diffusion models. In [9], the authors create a diffusion model taking into account protocol details such as the way of splitting files. However, the number of contaminated nodes with this model grows exponentially. The main contribution of this paper is the generalization of results observed in [7] and the explanation of this linear diffusion growth with an appropriate diffusion model.

This article is organized as follows: we describe in section 2 the network on which we study file diffusion in section 3. This diffusion has a linear behavior. In section 4, we propose a new diffusion model to reproduce this diffusion process. We interpret the results obtained with this model in section 5, before concluding and exposing our future work.

2. DATASET

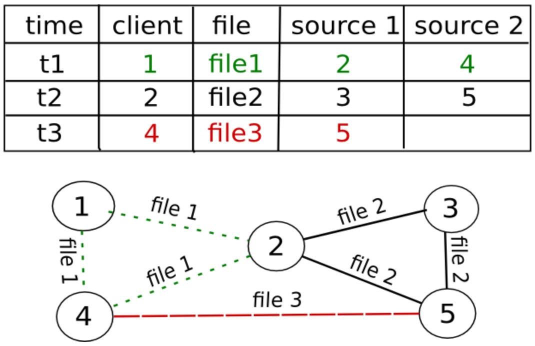
The dataset on which we study the diffusion process comes from a peer-to-peer network in which the activity of an eDonkey server was collected for two days [7]. The eDonkey protocol is a half-centralized peer-to-peer protocol: client peers make requests for some files, and the eDonkey server answers the client peer with potential source peers. These server responses have been collected, thus containing a set of sources for each given file, intended for a specific client. This client then retrieves the file from sources, which is not visible in traces because the server is no longer involved at this stage. Nodes session information is also available, indicating the periods of each client connection to the eDonkey server. We therefore know, at every moment, which nodes are present in the network, which represents the network's dynamics.

The measure lasted 48 hours. During this period, the server has received about 210 million requests for 2 million distinct files, and managed approximately 1.5 million client logins and logouts.

In order to simulate a diffusion on this dataset, we must first build the graph on which it occurs, in order to determine nodes neighbourhood. Indeed, in a diffusion model, an infected node may infect some of its neighbors. We build what we call the *interest graph*: when a client receives a server response, the server provides him with sources. Client and

[*]This work is supported in part by the French National Research Agency contract DynGraph ANR-10-JCJC-0202 and by the DiRe project, funded by the city of Paris *Émergence* program.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.



time	client	file	source 1	source 2
t1	1	file1	2	4
t2	2	file2	3	5
t3	4	file3	5	

Figure 1: Construction of the interest graph from 3 server responses.

sources are then interested in the same file. We consider that all of them correspond to connected nodes in the interest graph and form a clique, since they have a common interest in a file. Session information on nodes makes that graph dynamic: at a given time t, the interest graph only contains nodes logged to the server.
An example of interest graph built from server responses is given in figure 1: at time t1, the server indicates sources 2 and 4 to client 1 for file1. Peers 1, 2 and 4 are thus connected with one another in the interest graph. Note that a peer may be applicant (client) for some files, and supplier (source) for others. This is the case of peer 2, provider for file1 and applicant for file 2. Then, at time t2, the server indicates sources 3 and 5 to client 2 for file 2. Nodes 2, 3 and 5 are therefore connected with one another in the interest graph. We do the same at time t3, and we obtain the graph shown in figure 1.

3. REAL FILE DIFFUSION

We analyzed this peer-to-peer dataset to observe real files diffusion. In order to measure the average behavior of file diffusion, we took a dozen random samples of one thousand files among all files. Nodes which possess a given file are called ”infected” (with this specific file), whereas others are ”susceptible” to be infected, but are currently not. We then looked at the diffusion of each file, and calculated an average diffusion by looking at the average number of infected nodes at a time t. We observe (figure 2 top) that the number of infected nodes in the average real diffusion evolves linearly. This evolution is not usual, as in most works on diffusion on a peer-to-peer network, diffusions are modeled with an exponential growth rate [8] [9]. Typically, the SI model, frequently used for file diffusion, has an exponential behavior.

In order to validate this linear behavior, we compared it with the real spreading on another peer-to-peer dataset, collected during a longer time scale [1], which lasted ten weeks. Using the same method as for the short dataset, we found that the average behavior of the real spreading was linear too (figure 2 down), although its dynamics is very different.

The observation of nodes behaviors also showed that most clients do not share the file they just downloaded, i.e. they never become sources themselves. Indeed, the proportion of client nodes wich become sources is only equal to 8% in our dataset. In the following section, we first test a simple model inspired by SI, then we propose an improvement of this model, which better fits the actual linear spreading.

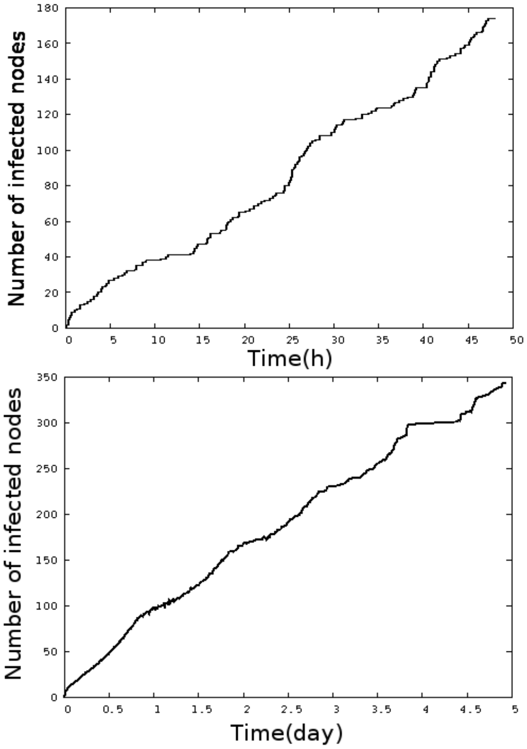


Figure 2: Real diffusion on both datasets. Top, short dataset. Bottom, long dataset.

4. MODELISATION

4.1 First model

Epidemiological models can easily be adapted to other contexts, in order to characterize many types of diffusion. The SI model is very simple, and used in many other cases than disease spreading. On our dataset, we first tested a diffusion model based on the SI model. Nodes can be in one of these two states: susceptible or infected. In the context of file diffusion, a node is susceptible if it does not have a certain file, and it is infected if it has this file. When a susceptible node is adjacent to an infected node, it has a certain probability p of being infected itself.

However, contamination in a peer-to-peer network is active, i.e. healthy nodes look for a file and download it. In our model, we represent this fact by keeping the same probability for a node of being infected, regardless of its number of infected neighbors. Thus, if we compare this model to a pure SI model, diffusion, although exponential, is slower since the contamination probability is invariant. This is where our model is different from the classical SI model.

Simulation results with this diffusion model are shown in figure 3 (bottom curve), compared to real diffusion (top curve). For this simulation, the probability p is equal to $\frac{1}{245000}$. We have chosen this parameter so that the simulation curve fitted the best the real diffusion curve.

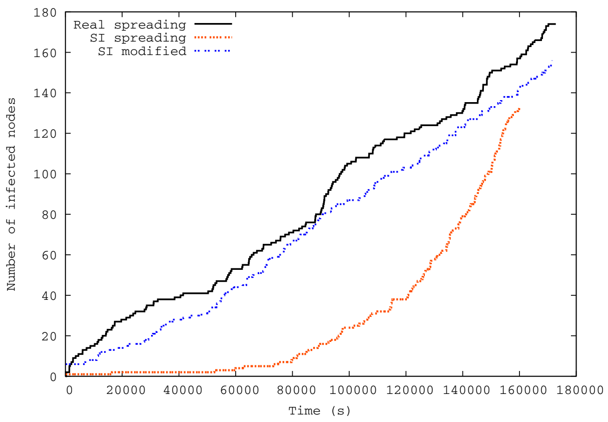


Figure 3: Real spreading (top), simulation of the first model (bottom) with $p = \frac{1}{30000}$, simulation of the second model (middle) with with $p = \frac{1}{30000}$ and $q = 0.08$.

We observe in our simulation that unlike real diffusion, the shape of the diffusion is clearly exponential. This result is not surprising with the model used: the more contaminated nodes, the higher the number of nodes in contact with infected nodes.

4.2 Second model

We then seeked the origin of this difference between real and simulated diffusion. The fact that most clients do not become sources is important, and should be taken into account: indeed, it significantly slows down file diffusion.
We therefore adapted the model proposed above to reflect this parameter and we introduced a new parameter: the probability that a node may infect other nodes (with probability p) once it is infected, i.e. its probability to be contagious. Two cases are thus possible when a node is infected: either it becomes contagious with probability q, or it is not contagious, in which case it cannot transmit the file at all.

This model has important differences with the SIR model (Susceptible / Infected / Removed), described in [6]. In the SIR model, nodes in the state I have a chance to "recover" and go to the state R: they are no longer counted as infected nodes. In the model proposed here, nodes that are contaminated and not contagious are still counted as infected nodes. Moreover, in the SIR model, a node in state I can go into the state R at every moment with a certain probability. In our model, blocking nodes, i.e. infected but not contagious for a given file, remain in this state until the end.

Looking into the data reveals that the proportion of clients who become providers is equal to 8%, so we simulated a diffusion with this new model, taking the probability that an infected node is contagious $q = 0.08$. Results of this simulation are shown in figure 3, for $p = \frac{1}{30000}$. To determine this parameter, we tested different values of p, while keeping the probability q constant, and we chose the value of p for which the simulation curve fitted the best the real diffusion curve. With this model, simulated diffusion follows a linear behavior, which is very close to the real diffusion. This model is therefore appropriate to modelize the type of diffusion that we observe here.

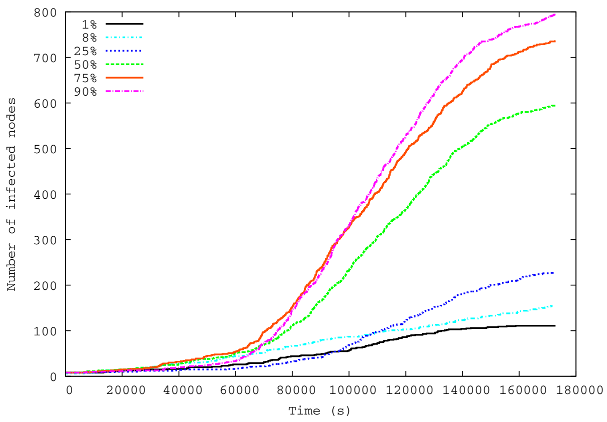


Figure 4: Diffusion simulation with model 2 with with $p = \frac{1}{30000}$ and with different values of q.

5. INTERPRETATION OF THE NEW MODEL

However, we want to explain how this model, yet inspired by SI that generally produces exponential curves, gives a linear growth in this case. The first model, proposed in section 4.1, shows an exponential behavior which is easily explained: when a node is infected, some of its neighbors become infected at time t. Then, at time t +1, nodes infected at step t will also spread the disease, and the number of contaminated will grow faster and faster. Why does the insertion of blocking nodes (nodes infected but not contagious) radically change the diffusion behavior?

5.1 Impact of the probability of being contagious

First, we want to observe the influence of the probability to become contagious when a node is contaminated. Indeed, if a contaminated node is contagious with 100% chance, the second model is identical to the first one.
In order to determine the impact of q parameter, we have performed several simulations of diffusion with different values of q. Results of these simulations are shown in figure 4. Parameters used for these different simulations correspond to probabilities of becoming contagious equal to 0.01, 0.08, 0.25, 0.5, 0.75 and 0.9. We observe that with the lowest values of q, diffusion behavior is linear, like in the case of model 2 with a parameter of 0.08. However, when q increases, diffusion has an exponential behavior. When the probability is very high, this model is very similar to the first model.

5.2 Influence of graph structure

At each time slot, all nodes are contaminable, i.e. have at least one contagious neighbor. The diffusion therefore only depends on the value of p. A node infected at time t, with N neighbors (N very large compared to 1 in our graph) infects P of its neighbors (P is $o(N)$ because the probability of contamination is low). So at time t + 1, P nodes are infected, and only a small part become contagious. This number is very small compared to N and therefore to the number of potentially contaminable nodes. By repeating this process, we thus obtain a linear behavior. So in the case of peer-to-peer eDonkey networks, diffusion is somehow independent of the graph structure, since all nodes can be reached at each

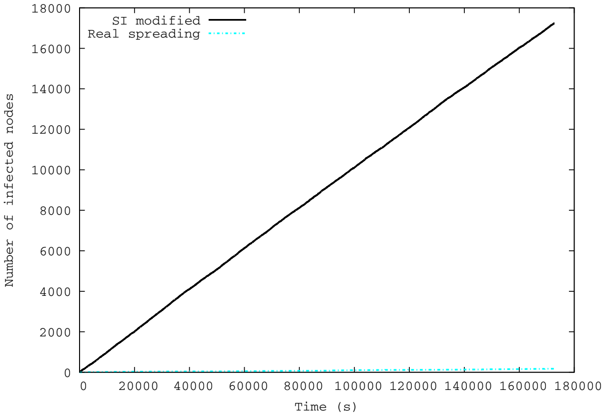


Figure 5: Diffusion with the second model on the interest graph, without considering nodes dynamics (top curve), with $p = \frac{1}{30000}$ and $q = 0.08$ and real spreading (bottom curve).

time. Only the dynamics of the graph influence the diffusion behavior, by changing the set of nodes present at each time slot (and therefore potentially contaminable).

In order to estimate the impact of dynamics on the diffusion behavior, we simulated a diffusion using model 2 on the interest graph, without taking into account information of dynamics about nodes: we consider the whole graph at every moment. The results of this simulation are shown in figure 5. We observe that the diffusion behavior is still linear; however, the number of infected nodes grows much faster than if we consider the dynamic graph.

6. CONCLUSION AND FUTURE WORK

In this work, we have observed a file diffusion phenomenon on a peer-to-peer network. The observed diffusion has a linear behavior, while in works already done on this topic, diffusion of exponential type are studied. In order to simulate diffusion on this network, we had to make hypotheses, like for example, deciding how to construct the interest graph. We proposed a model, inspired by the SI model, which presents a linear behavior, in order to approximate correctly the real spreading. We also explained why this model is linear, and we studied the behavior of this model by varying its parameter values.

In the future, we will first study diffusion on other types of peer-to-peer networks: eDonkey is a half-centralized peer-to-peer protocol, with a specific type of dynamics. Different peer-to-peer protocols, such as BitTorrent or Gnutella may present an other diffusion behavior, and nodes dynamics can be very different. Under these circumstances, it would be interesting to know if our model is consistent whith other types of peer-to-peer networks.

In a longer term, we will also determine for what other types of diffusion this model may be appropriate. In the peer-to-peer network we have studied in this paper, diffusion is active, i.e. nodes choose to download from suppliers a file that interests them . This behavior may be related to adoption models. It would thus be interesting to apply the diffusion model proposed here in the case of innovation diffusion.

7. REFERENCES

[1] F. Aidouni, M. Latapy, and C. Magnien. Ten weeks in the life of an edonkey server. *Hot P2P*, 2009.

[2] S. Aral, E. Brynjolfsson, and M. V. Alstyne. Productivity effects of information diffusion in networks. 2007.

[3] L. Cabral. 15. Equilibrium, epidemic and catastrophe: diffusion of innovations with network effects. 2002.

[4] S. Cauchemez, A. Bhattarai, T. L. Marchbanks, R. P.Fagan, S. Ostroff, N. M.Ferguson, D. Swerdlow, and the Pennsylvania H1N1 working group. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *PNAS*, 2010.

[5] F. Deroïan. Formation of social networks and diffusion of innovations. *Research Policy*, 31:835–846, 2002.

[6] M. Girvan, D. S. Callaway, M. Newman, and S. H.Strogatz. A simple model of epidemics with pathogen mutations. *Physical Review*, 2002.

[7] J.-L. Guillaume, M. Latapy, and S. Le-Blond. Statistical analysis of a p2p query graph based on degrees and their time-evolution. *IWDC*, 2004.

[8] P. Han, K. Hosanager, and Y.-W. Tan. Diffusion of digital products in peer-to-peer networks. 2004.

[9] K. Leibnitz, T. Hoßfeld, N. Wakamiya, and M. Murata. Modeling of epidemic diffusion in peer-to-peer file-sharing networks. 2006.

[10] F. Pasquale. Information Spreading in Dynamic Networks. *Networks*, 2008.

[11] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *Social Networks*, pages 1–4, 2010.

[12] G. Serazzi and S. Zanero. Computer virus propagation models. *LNCS*, 2004.

[13] J. Sthelé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J. F. Pinton, N. Khanafer, W. V. den Broeck, and P. Vanhems. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, 2011.

[14] T. Zhou, J.-G. Liu, W.-J. Bai, G. Chen, and B.-H. Wang. Behavior of susceptible-infected epidemics on scale-free networks with identical infectivity. *Physical review*, 2006.