

Information dissemination processes in directed social networks

K. Avrachenkov^{*1}, K. De Turck^{†2}, D. Fiems^{‡2}, and B.J. Prabhu^{§3,4}

¹INRIA Sophia Antipolis, 2004 route des Lucioles, 06902 Sophia Antipolis Cedex, France

²Department of Telecommunications and Information Processing, Ghent University, St. Pietersnieuwstraat 41, 9000 Gent, Belgium

³CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France

⁴Univ de Toulouse, LAAS, F-31400 Toulouse, France

November 11, 2013

Abstract

Social networks can have asymmetric relationships. In the online social network Twitter, a follower receives tweets from a followed person but the followed person is not obliged to subscribe to the channel of the follower. Thus, it is natural to consider the dissemination of information in directed networks. In this work we use the mean-field approach to derive differential equations that describe the dissemination of information in a social network with asymmetric relationships. In particular, our model reflects the impact of the degree distribution on the information propagation process. We further show that for an important subclass of our model, the differential equations can be solved analytically.

1 Introduction

We develop mathematical models for the dissemination of information on directed graphs and investigate the influence of parameters such as the degree distribution on the dynamics of the dissemination process. The directed graph model, as opposed to the undirected model, is better suited for networks like the one of Twitter because of the asymmetric relationship that exists between

^{*}k.avrachenkov@sophia.inria.fr

[†]kdeturck@telin.ugent.be

[‡]Dieter.Fiems@telin.UGent.be

[§]balakrishna.prabhu@laas.fr

different users. Specifically, in the Twitter network, a user can choose to receive the tweets—in other words, become a follower—of one or more other users by subscribing to their accounts. Certain users, celebrities for example, have several millions of followers who follow their tweets. These users do not necessarily follow the tweets of all of their followers which results in an asymmetric relationship between users. This asymmetry is modelled by a directed graph in which an outgoing edge is drawn from a user to each of its followers. An edge in the opposite direction from the follower to the user need not always exist and is drawn only if the user subscribes to the channel of this follower.

A hashtag is a word or a phrase prefixed by # and used in social networks as a keyword. The prefix facilitates the search for conversations related to the prefixed word or phrase. The typical life cycle of a hashtag closely resembles an epidemic. In the first phase the interest in the hashtag grows as users generate tweets containing this hashtag. These tweets are received by followers who then either retweet them or generate new tweets with this hashtag. The number of users tweeting this hashtag (“infected” users) grows as a function of time during this phase. At a certain point in time, the interest reaches its zenith and starts to wane as users move on and get interested in other events. The second phase begins at this point in time as users stop tweeting this hashtag (or, “recover”), and the number of infected users decreases.

In this work we use the mean-field approach to derive the differential equations which describe the process of information dissemination. We obtain a couple of differential equations which describe the evolution of the fractions of infected and recovered persons. As a model for the underlying network we take the Configuration-type model for directed graphs [2]. While epidemics have been widely studied on undirected graphs, there is hardly any analysis of the epidemic-type processes on directed networks. In [7, 6], the mean-field approach has been applied to the analysis of epidemics on an undirected configuration-type graph model. In [5], the effect of network topology has been analysed in the case of undirected graphs. In particular, the authors of [5] applied their general results to analyse the Erdős-Rényi and preferential attachment random graph models. An interesting approach combining a decomposition approach with two-state primitive Markov chain has been proposed in [8] for undirected networks with general topology. For an overview of various results about epidemic processes on undirected networks we refer the interested reader to the books [4, 1, 3].

2 The mean-field model

Consider a network of N nodes structured according to the Configuration-type model for directed graphs [2]. The in-degree and out-degree of the nodes are drawn from a distribution $f(k, l) = \mathbb{P}(K = k, L = l)$, defined on the bounded set $\mathcal{D} = \{(k, l) : 0 \leq k \leq \hat{K}, 0 \leq l \leq \hat{L}, (k, l) \neq (0, 0)\}$, where the first (resp. second) index corresponds to the in-degree (resp. out-degree) and \hat{K} (resp. \hat{L}) is the maximal in-degree (resp. out-degree). In the remainder, we always assume

that $\mathbb{E}K = \mathbb{E}L$; in a network every outgoing link is an incoming link of some other node. A generic node with in-degree k and out-degree l shall be referred to as a (k, l) -node.

Each node in the network can be in one of the three states : infected, recovered, or susceptible. An infected node infects its susceptible neighbours after an exponentially distributed time with intensity λ . Note that, the infection is spread simultaneously along all the outgoing edges and not just one edge at a time. The simultaneous dissemination along all outgoing edges models the spread of tweets on Twitter. An infected node recovers after an exponentially distributed time of rate ν , at which time it stops spreading information in the network.

We shall be mainly interested in a large-population model, that is when $N \rightarrow \infty$. This assumption simplifies considerably the analysis of the dissemination process while being realistic¹.

Let $i_{k,l}(t)$ (resp. $r_{k,l}(t)$) denote the fraction of infected (resp. recovered) (k, l) nodes at time t . The following result describes the dynamics of these two quantities.

Theorem 1. *Let $i_{k,l}(0) > 0$ for some $(k, l) \in \mathcal{D}$. Then, $\forall (k, l) \in \mathcal{D}$,*

$$\frac{di_{k,l}(t)}{dt} = \lambda k(f(k, l) - i_{k,l}(t) - r_{k,l}(t)) \frac{\sum_{k',l'} l' i_{k',l'}(t)}{\sum_{k',l'} l' f(k', l')} - i_{k,l}(t)\nu, \quad (1)$$

and

$$\frac{dr_{k,l}(t)}{dt} = i_{k,l}(t)\nu. \quad (2)$$

Sketch of proof. Let $I_{k,l}^N(t)$ (resp. $R_{k,l}^N(t)$) be the number of infected (resp. recovered) (k, l) nodes in a network of N nodes. Then in a small time interval Δ ,

$$\begin{aligned} I_{k,l}^{(N)}(t + \Delta) &= I_{k,l}^{(N)}(t) + \text{number of } (k, l) \text{ nodes infected in time } \Delta \\ &\quad - \text{number of } (k, l) \text{ infected } (k, l) \text{ nodes that recover in time } \Delta. \end{aligned}$$

Since each infected node recovers after an exponentially distributed time of rate ν , the number of (k, l) infected nodes that recover in Δ will be approximately $I_{k,l}^{(N)}(t)\nu\Delta$. There will be additional terms containing Δ^2 which we neglect.

Let us compute the number of (k, l) nodes that get infected in time Δ . There are $N_{k,l}^{(N)} - (I_{k,l}^{(N)}(t) + R_{k,l}^{(N)}(t))$ susceptible (k, l) nodes. Assume that each (k, l) node has a probability $p_{k,l}$ to get infected in the interval Δ . Then, expected number of infected (k, l) nodes in time Δ will be $(N_{k,l}^{(N)} - (I_{k,l}^{(N)}(t) + R_{k,l}^{(N)}(t)))p_{k,l}$.

Each (k, l) node has k incoming edges. Assuming that the edges are connected independently, $p_{k,l} = (1 - (1 - q_{k,l})^k)$, where $q_{k,l}$ is the probability that the infection is transmitted along one of the edges in Δ . The number of (k, l) nodes infected in Δ is thus

$$(N_{k,l}^{(N)} - (I_{k,l}^{(N)}(t) + R_{k,l}^{(N)}(t))) \cdot (1 - (1 - q_{k,l})^k),$$

¹Twitter has approximately 200 million registered users (source: Wikipedia).

which, for Δ sufficiently small, can be approximated as

$$(N_{k,l}^{(N)} - (I_{k,l}^{(N)} + R_{k,l}^{(N)}))kq_{k,l}.$$

Finally, we compute $q_{k,l}$. In an interval Δ , each infected node transmits the infection with probability $\lambda\Delta$. Thus, there are $\sum_{k',l'} l' I_{k',l'}^{(N)} \lambda\Delta$ edges that are infected and that transmit the infection in Δ . There are a total of $\sum_{k',l'} l' N_{k',l'}^{(N)}$. Assuming that an incoming edge is connected uniformly at random to an outgoing edge, we obtain

$$q_{k,l} = \frac{\sum_{k',l'} l' I_{k',l'}^{(N)} \lambda\Delta}{\sum_{k',l'} l' N_{k',l'}^{(N)}}.$$

Consequently,

$$I_{k,l}^{(N)}(t + \Delta) - I_{k,l}^{(N)}(t) = (N_{k,l}^{(N)} - (I_{k,l}^{(N)} + R_{k,l}^{(N)}))k \frac{\sum_{k',l'} l' I_{k',l'}^{(N)} \lambda\Delta}{\sum_{k',l'} l' N_{k',l'}^{(N)}} - I_{k,l}^{(N)} \nu \Delta.$$

If the initial number of nodes is large, then we can divide the two sides of the above equation to obtain the following difference equation in terms of the fraction of the infected and the recovered nodes:

$$i_{k,l}(t + \Delta) - i_{k,l}(t) = (f(k, l) - (i_{k,l}(t) + r_{k,l}(t)))k \frac{\sum_{k',l'} l' i_{k',l'}(t) \lambda\Delta}{\sum_{k',l'} l' f(k', l')} - i_{k,l}(t) \nu \Delta.$$

To complete the picture, we take the limit $\Delta \rightarrow 0$, and obtain the differential equations (1) and (2). \square \square

Remark 1. In the above equation $i_{k,l}$ is the fraction of the (k, l) nodes that are infected. This fraction varies between 0 and $f(k, l)$. If instead, we want to look at the evolution of the fraction of infected nodes and recovered nodes conditioned on them being (k, l) nodes, then the corresponding differential equations for these fractions will be

$$\frac{di_{k,l}(t)}{dt} = \lambda k (1 - i_{k,l}(t) - r_{k,l}(t)) \frac{\sum_{k',l'} l' f(k', l') i_{k',l'}(t)}{\sum_{k',l'} l' f(k', l')} - i_{k,l}(t) \nu, \quad (3)$$

$$\frac{dr_{k,l}(t)}{dt} = i_{k,l}(t) \nu. \quad (4)$$

3 Epidemics without recovery

The solution of (1) and (2) can be computed numerically. In some specific case we can obtain explicit solutions to these equations. In particular, this is the case when there is no recovery: $\nu = 0$, or in the language of Twitter, they keep generating new tweets with the same hashtag. That is, a hashtag never gets out of mode. This can well represent the case for the topics or personalities that can sustain popularity over a long period of time.

Since there are no recovered nodes, $r_{k,l}(t) = 0, \forall t$, and (1) takes the form

$$\frac{di_{k,l}(t)}{dt} = \lambda k(f(k,l) - i_{k,l}(t)) \frac{\sum_{k',l'} l' i_{k',l'}(t)}{\sum_{k',l'} l' f(k',l')}. \quad (5)$$

The differential equation (5) can be solved in terms of a reference value of (k,l) , say $(k,l) = (1,1)$ by noting that

$$\frac{1}{k(f(k,l) - i_{k,l}(t))} \frac{di_{k,l}(t)}{dt} = \frac{1}{(f(1,1) - i_{1,1}(t))} \frac{di_{1,1}(t)}{dt},$$

whence

$$f(k,l) - i_{k,l}(t) = \frac{f(k,l) - i_{k,l}(0)}{(f(1,1) - i_{1,1}(0))^k} (f(1,1) - i_{1,1}(t))^k =: c_{k,l} (f(1,1) - i_{1,1}(t))^k. \quad (6)$$

The fraction of infected nodes of degree $(1,1)$ can be obtained by substituting the value of $i_{k,l}(t)$ in (5) and solving it:

$$\frac{di_{1,1}(t)}{dt} = \lambda(f(1,1) - i_{1,1}(t)) \frac{\sum_{k',l'} l' (f(k',l') - c_{k',l'} (f(1,1) - i_{1,1}(t))^{k'})}{\sum_{k',l'} l' f(k',l')}. \quad (7)$$

Deterministic in-degree

Assume that the in-degree K is deterministic and is equal to d . Then, equation (5) becomes

$$\frac{di_{d,l}(t)}{dt} = \lambda d(f(d,l) - i_{d,l}(t)) \sum_{l'} \frac{l' i_{d,l'}(t)}{\sum_j j f(d,j)}.$$

Since the expected in-degree and the expected out-degree coincide, $\sum_j j f(d,j) = d$. Denote $\Theta(t) = \sum_j \frac{j i_{d,j}(t)}{d}$, and rewrite the above equation as:

$$\frac{di_{d,l}}{dt} = \lambda d(f(d,l) - i_{d,l}(t)) \Theta(t). \quad (8)$$

Multiplying the above equation by $\frac{l}{d}$ and summing over all values l , we obtain the following equation for Θ :

$$\frac{d\Theta}{dt} = \lambda d(1 - \Theta)\Theta,$$

which upon integration yields:

$$i_{d,l}(t) = f(d,l) - c_{1e} e^{-\lambda d \int \Theta(t) dt} = f(d,l) - \frac{f(d,l) - i_{d,l}(0)}{1 - \Theta(0) + \Theta(0)e^{-\lambda d t}}. \quad (9)$$

4 Numerical experiments

In this section, we validate the mean-field model developed in Section 2. In the numerical experiments, first the in-degree and out-degree sequences are generated according to the given degree distributions. So as to have the same number of incoming stubs as outgoing stubs, the difference between the two is added to the smaller quantity. A configuration-type graph is then created by matching an incoming stub with an outgoing stub chosen uniformly at random. It was shown in [2] that this procedure does indeed approximate closely the configuration model. The information diffusion process is then simulated on this graph.

For computing the solution of the system of differential equations (1) and (2) numerically, the empirical degree distributions from the graph generated previously are given as input.

The results of two such experiments with 20000 nodes is shown in Figure 1. The in-degree and the out-degree sequences were taken to be independent of each other. The out-degree sequence was drawn from a Uniform distribution in the set $\{1, 20\}$ in the two simulations. For the figure on the left, the in-degree distribution was taken to be deterministic with parameter 10, and for the figure on the right it was the Zipf law on $\{1, 71\}$ and exponent 1.2. In both experiments, $\lambda = 1$ and $\nu = 0.5$, and 5 percent of all nodes were assumed to be infected at time 0.

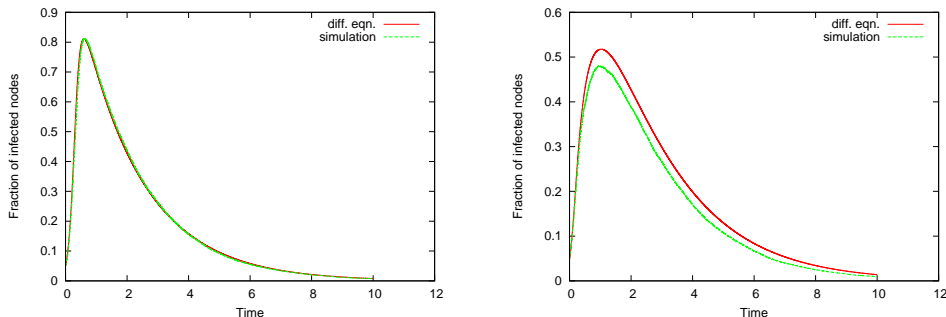


Figure 1: Fraction of all nodes infected as a function of time for Deterministic in-degree distribution (left) and Zipf in-degree distribution (right).

Observations

It is observed that the dissemination process is faster when the variance of the in-degree distribution is smaller. This observation was reinforced by other experiments in which the in-degree was drawn from a Uniform distribution. In several other experiments that we conducted, it was also observed that the out-degree distribution does not have any noticeable effect of the dynamics of the epidemics.

Our on-going work is oriented towards investigating the influence of the variance of the in-degree distribution and giving a theoretical foundation to the above observations.

5 Acknowledgments

This work was partially supported by the Partenariat Hubert Curien PHC Tournesol FR 2013 29053SF between France and the Flemish community of Belgium, by the Inria Alcatel-Lucent Joint Lab ARC “Network Science”, and by the European Commission within the framework of the CONGAS project FP7-ICT-2011-8-317672.

References

- [1] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.
- [2] N. Chen and M. Olvera-Cravioto. Directed random graphs with given degree distributions. *To appear in Stochastic Systems*.
- [3] Moez Draief and Laurent Massouli. *Epidemics and Rumours in Complex Networks*. Cambridge University Press, 2010.
- [4] Richard Durrett. *Random Graph Dynamics*, volume 20. Cambridge university press, 2007.
- [5] Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 2, pages 1455–1466. IEEE, 2005.
- [6] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 26(4):521–529, 2002.
- [7] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.
- [8] P. Van Mieghem, J. Omic, and R. Kooij. Virus spread in networks. *Networking, IEEE/ACM Transactions on*, 17(1):1–14, 2009.