Supplementary Information
Activity driven modeling of time varying networks

N. Perra, B. Gonçalves, R. Pastor-Satorras, A. Vespignani

May 11, 2012

Contents

1 The Model **1**
1.1 Integrated network 2

2 Epidemic threshold **4**
2.1 Case $m = 1$ 4
2.2 Case $m > 1$ 5

3 Persistence of links **5**

4 Dataset Details **6**
4.1 PRL Dataset 6
4.2 Twitter Dataset 8
4.3 IMDb Dataset 8

1 The Model

Let us consider N initially disconnected individuals/nodes. We define for each agent i the activity potential x_i as the probability that any social act/connection in the system has been engaged by the actor i. We define $F(x)$ as the probability distribution that a randomly chosen agent, i, has activity potential x. To avoid possible divergences of the distribution close to 0 we fix a bound by imposing $x \in [\epsilon, 1]$. Given the activity potential x_i is natural to define the activity rate a_i as:

$$a_i = \eta x_i, \tag{1}$$

where η is a rescaling factor defined such that the average number of active nodes per unit time in the system is $\eta\langle x\rangle N$. At each time step t the network $\mathcal{G}_t$ is build starting from N disconnected vertices (all the edges in the network are deleted). The links are created in the following way:

- With probability $a_i \Delta t$ the vertex i becomes active (*fires*) and generates m links that are connected to m other vertex selected randomly

- With probability $1 - a_i \Delta t$ the node will not fire, but can however receive connections from other active vertices.

1.1 Integrated network

We define the integrated network $\mathcal{G}$ as the union of all networks obtained in each time step,

$$\mathcal{G} = \bigcup_{t=0}^{t=T} \mathcal{G}_t. \tag{2}$$

Multiple edges or self-links are not allowed. For a given distribution of activity potential, $F(x)$, we are interested in computing the degree distribution of the integrated network at time T, $P_T(k)$. Let us consider the number $\tau = T/\Delta t$ of instantaneous networks generated up to time T. The number of times that the vertex i will be active is given by a binomial distribution with an average $\tau a_i \Delta t = T a_i$. At time T the average number of active nodes will be

$$\sum_i T a_i = TN\langle a \rangle, \tag{3}$$

while in the single instantaneous network will be

$$N_t = N\langle a \rangle \Delta t \equiv N\eta \langle x \rangle \Delta t. \tag{4}$$

Each active node will create m links. Therefore, the average number of edges per unit time will be on average:

$$E_t = mN\eta\langle x \rangle \tag{5}$$

leading to an average degree per unit time

$$\langle k \rangle_t = \frac{2E_t}{N} = 2m\eta\langle x \rangle. \tag{6}$$

The instantaneous network will be composed by a set of stars, the vertices that were active at that time step, with degree larger than or equal to m, plus some vertices with low degree. The structure of the integrated network will be however much complex. Consider this integrated network at time T. The degree of each one of its nodes can be written as $k_T(i) = k_T^{out}(i) + k_T^{in}(i)$, where the out-degree $k_T^{out}(i)$ corresponds to the links emanating from i due to its becoming active, while the in-degree $k_T^{in}(i)$ is due to the links that have arrived to i from other active nodes. Let us focus on the out-degree. In the time interval T, node i will have tried to send in average Tma_i edges. Not all those edges will contribute to its integrated out-degree, just only those arriving to different nodes (no multiple links are allowed in the integrated network). The out-degree can thus be computed by making an analogy to the Polya urns problem: it will be equal to the number of different balls extracted from a urn with N balls, performing Tma_i extractions. The probability of extracting d balls will be given by

$$P(d) = \binom{N}{d} p^d (1-p)^{(N-d)} \tag{7}$$

where

$$p = 1 - \left(1 - \frac{1}{N}\right)^{Tma_i} \tag{8}$$

is the probability of extracting at least one ball in the urn. The average value is then

$$\langle d \rangle = pN. \tag{9}$$

Therefore the average out-degree of a vertex i in the integrated network can be estimated as

$$k_T^{out}(i) = N(1 - e^{-Tma_i/N}) \tag{10}$$

in the limit of large N and small T/N.

The in-degree will come from the rest of vertices that are active and have sent connections to i, without receiving them from it. In T time steps, vertex i will have fired Ta_i times in average. The probability that a vertex has not received any connection form i will then be

$$\left(1 - \frac{1}{N}\right)^{mTa_i} \simeq \exp[-mTa_i/N], \tag{11}$$

where we assume again large N and small T/N. The average number of vertices that have fired and are not connected to i with and egde that emanated from i will then be $TN\langle a \rangle \exp[-mTa_i/N]$. These nodes have m possibilities to reach at random i, each with probability $1/N$. The number of connections that have reached this node will then be on average $mT\eta\langle x \rangle \exp[-mTa_i/N]$. The degree of the node i will be simply the sum of these to contributions:

$$\begin{aligned} k_T(i) = k_T^{out}(i) + k_T^{in}(i) \quad &= N(1 - e^{-Tma_i/N}) + Tm\eta\langle x \rangle e^{-Tma_i/N} \qquad (12) \\ &= N\left[1 - \left(1 - m\eta\langle x \rangle \frac{T}{N}\right) e^{-Tma_i/N})\right] \\ &\sim N(1 - e^{-Tma_i/N}) = N(1 - e^{-Tm\eta x_i/N}), \qquad (13) \end{aligned}$$

again for small T/N. From the last relation we can write the activity potential x as an effective function of the integrated degree k, i.e.:

$$x(k) = -\frac{N}{\eta m T} \ln\left(1 - \frac{k}{N}\right). \tag{14}$$

At the probabilistic level we can use the relation $P_T(k)dk \sim F(x)dx$, where P is the degree distribution of the integrated, network after T time steps, to finally obtain

$$P_T(k) \sim F[x(k)]\frac{dx(k)}{dk} = \frac{1}{Tm\eta}\frac{1}{1 - \frac{k}{N}} F\left[-\frac{N}{\eta m T} \ln\left(1 - \frac{k}{N}\right)\right]. \tag{15}$$

In the limit of small k/N (for not very large values of T) we can expand the logarithm, to obtain the simplified expression

$$P_T(k) \sim \frac{1}{Tm\eta} F\left[\frac{k}{Tm\eta}\right]. \tag{16}$$

That is, we obtain an important relation binding the functional form of the degree distribution to the activity distribution of the nodes.

2 Epidemic threshold

Let us consider the SIS epidemic compartmental model, characterized by a transition probability λ and a recovery time μ^{-1}, spreading in the dynamical network generated as discussed above. Let us assume a distribution of activity a of nodes given by a general distribution $F(x)$ as before. At a mean-field level, the epidemic process will be characterized by the number of infected individuals in the class of activity a, at time t, namely I_a^t.

2.1 Case $m = 1$

The number of infected individuals of class a at time $t + \Delta t$ given by:

$$I_a^{t+\Delta t} = -\mu \Delta t I_a^t + I_a^t + \lambda(N_a^t - I_a^t) a \Delta t \int da' \frac{I_{a'}^t}{N} + \lambda(N_a^t - I_a^t) \int da' \frac{I_{a'}^t a' \Delta t}{N}, \tag{17}$$

where N_a is the total number of individuals with activity a. In Eq. (17), the third term on the right side takes into account the probability the a susceptible of class a is active and get the infection getting a connection from any other infected individual (summing over all different classes), while the last term takes into account the probability that a susceptible, independently of his activity, gets a connection from any infected active individual. Now summing on all the classes we get (ignoring the second order terms):

$$\int da I_a^{t+\Delta t} = I^{t+\Delta t} = I^t - \mu \Delta t I^t + \lambda \langle a \rangle I^t \Delta t + \lambda \theta^t \Delta t, \tag{18}$$

where $\theta^t = \int da' I_{a'}^t a'$. We can get another expression multiplying both sides of Eq. (17) by a and integrating, to obtain

$$\theta^{t+\Delta t} = \theta^t - \mu \theta^t \Delta t + \lambda \langle a^2 \rangle I^t \Delta t + \lambda \langle a \rangle \theta^t \Delta t. \tag{19}$$

In the limit $\Delta t \to 0$, we can write Eqs. (17) and (19) in a differential form:

$$\partial_t I = -\mu I + \lambda \langle a \rangle I + \lambda \theta, \tag{20}$$

$$\partial_t \theta = -\mu \theta + \lambda \langle a^2 \rangle I + \lambda \langle a \rangle \theta. \tag{21}$$

The Jacobian matrix of this set of linear differential equations takes the form

$$J = \begin{pmatrix} -\mu + \lambda \langle a \rangle & \lambda \\ \lambda \langle a^2 \rangle & -\mu + \lambda \langle a \rangle \end{pmatrix},$$

and has eigenvalues

$$\Lambda_{(1,2)} = \lambda \langle a \rangle - \mu \pm \lambda \sqrt{\langle a^2 \rangle}. \tag{22}$$

The epidemic threshold is obtained requiring the largest eigenvalues to be larger the 0, which leads to the condition for the presence of an endemic state:

$$\frac{\lambda}{\mu} > \frac{1}{\langle a \rangle + \sqrt{\langle a^2 \rangle}} + \mathcal{O}(\frac{1}{N}) \tag{23}$$

The order $1/N$ is present because we are not considering events in which two infected nodes choose each other for connection.

An important epidemiological quantity is the reproductive number R_0, defined as the average number of secondary cases generated by a primary case in an entirely susceptible population. For a SIS model we have

$$R_0 = \frac{\beta}{\mu} \equiv \frac{\lambda\langle k\rangle}{\mu} \tag{24}$$

where $\beta = \lambda\langle k\rangle$ is the per capita spreading rates that takes into account the rate of contacts of each individual. Considering this in the equation (23) at the first order we get as a threshold in terms of the reproductive numner:

$$R_0 > R_0^C = \frac{2\langle a\rangle}{\langle a\rangle + \sqrt{\langle a^2\rangle}} = \frac{2\langle x\rangle}{\langle x\rangle + \sqrt{\langle x^2\rangle}} = \frac{2}{1 + \sqrt{\frac{\langle x^2\rangle}{\langle x\rangle^2}}} \tag{25}$$

2.2 Case $m > 1$

In this general case we can use the same machinery as above. We have just to add the fact that each active agent can now make more then one connection (m at most) at each trial. In this case the probability to get a connection changes:

$$\frac{1}{N} \to 1 - \left(1 - \frac{1}{N}\right)^m \sim \frac{m}{N} \tag{26}$$

Using this fact, we can work out the corresponding set of differential equations, whose Jacobian matrix is now:

$$J_m = \begin{pmatrix} -\mu + \lambda m\langle a\rangle & \lambda m \\ \lambda m\langle a^2\rangle & -\mu + \lambda m\langle a\rangle \end{pmatrix},$$

with eigenvalues

$$\Lambda_{(1,2)} = m\lambda\langle a\rangle - \mu \pm \lambda m\sqrt{\langle a^2\rangle}. \tag{27}$$

That is, the eigenvalues in the case $m > 1$. Thus, the epidemic threshold in the general case can be simply written as

$$R_0 > R_0^C = \frac{1}{m}\frac{2m\langle a\rangle}{\langle a\rangle + \sqrt{\langle a^2\rangle}} = \frac{2\langle x\rangle}{\langle x\rangle + \sqrt{\langle x^2\rangle}} \tag{28}$$

3 Persistence of links

In the model, at each time step a random markovian network is created. There is no memory. Each link created at time step $t - \Delta t$ is deleted and recreated randomly at time step t. This is an oversimplification of real social interaction. In the real world we can imagine that individuals establish social connections preferentially within an limited circle of friends, and that they have a memory. This implies that a certain number of connections are active for more than one single time window Δt, they are persistent. In Figure S1 we show the distribution of persistence for the APS dataset, in Figure S2 for Twitter and in Figure S3 for the IMDb. As it becomes clear from the plots,

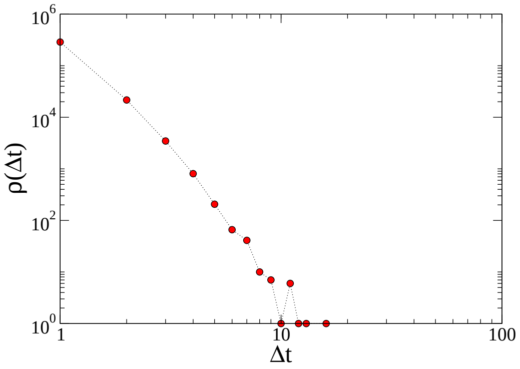

Figure S1: For the APS database we plot the distribution of connections active for Δt consecutive time steps: persistence of links. In this case one time window correspond to one year

a relevant number of connections stay active across different time windows. In our model we consider instead a mean-field approach where temporal correlations between users are neglected. However it is easy to generalize our model in order to take into account the persistence of links and its effect in the spreading of an epidemic disease. This will be a matter of future work.

4 Dataset Details

We study three different dataset: the collaborations in the journal "Physical Review Letters" (PRL) published by the APS[1], the message exchanged on Twitter and the activity of actors in movies and TV series as recorded in the Internet Movie Database (IMDb)[2].

4.1 PRL Dataset

In this database the network representation considers each author of a PRL article as a node. An undirected link between two different authors is drawn if they collaborated in the same article. We filter out all the articles with more than 10 authors in order to focus our attention just on small collaborations in which we can assume that the social components is relevant. We consider the period between 1960 and 2004. In this time window we registered 71.583 active nodes and 261.553 connections among them.

In this dataset is natural defining the activity rate, a, of each author as the number of papers written in a specific time window Δt.

[1] The data are available here: `http://prx.aps.org/node/3966`

[2] The data are available here: `http://www.imdb.com/interfaces`

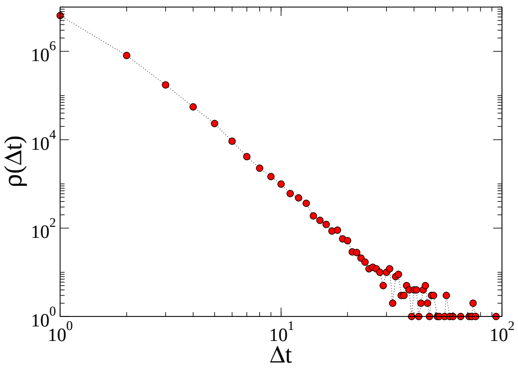

Figure S2: For the Twitter database we plot the distribution of connections active for Δt consecutive time steps: persistence of links. In this case one time window correspond to one day

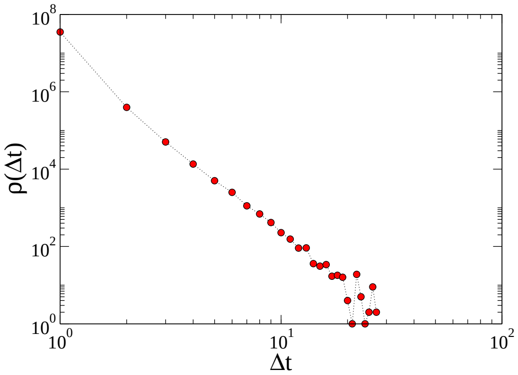

Figure S3: For IMDb we plot the distribution of connections active for Δt consecutive time steps: persistence of links. In this case one time window correspond to one year.

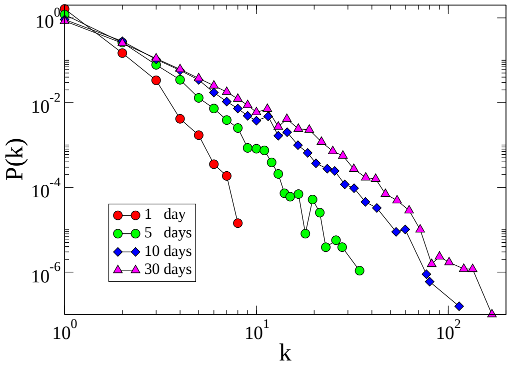

Figure S4: Considering the subset of users active on the first 31 days of our twitter dataset we plot the degree distribution of the resulting networks for different integrating time windows.

4.2 Twitter Dataset

Having been granted temporary access to Twitter's firehose we mined the stream for over 6 months to identify a large sample of active user accounts. Using the API, we then queried for the complete history of 3 million users, resulting in a total of over 380 million individual tweets covering almost 4 years of user activity on Twitter. In this database the network representation considers each users as a node. An undirected link between two different users is drawn if they exchanged at least one message. We focus our attention on 9 months during 2008. In this time window we registered 531.788 active nodes and 2.566.398 connections among them.

In this dataset we define the activity rate of each user as the number of messages sent in a time window Δt. In figure S4 we show the degree distribution of the subgraph obtained by integrating over $1, 5, 10, 30$ days. The subset of nodes considered are those active in the the first months of the dataset (January 2008). It is clear how the network integrated in one day has a sparse nature. Increasing the time window heterogeneous connectivity patterns start to emerge.

4.3 IMDb Dataset

In this database the network representation considers each actor as a node. An undirected link between two different actors is drawn if they collaborated in the same movie/TV series. We focus on the period between 1950 and 2010. During this time period we registered 1.273.631 active nodes and 47.884.882 connections between them.

A natural way to define the activity rate in this dataset is to consider the number of movies acted by each actor in a specific time window Δt. In figure S5 we show the degree distribution of the subgraph obtained by integrating over $1, 5, 10, 30$ years. The subset of actors is formed by those active in the period $1970 - 1980$. Even in this case is clear how increasing the integrating time window the level of heterogeneity increases.

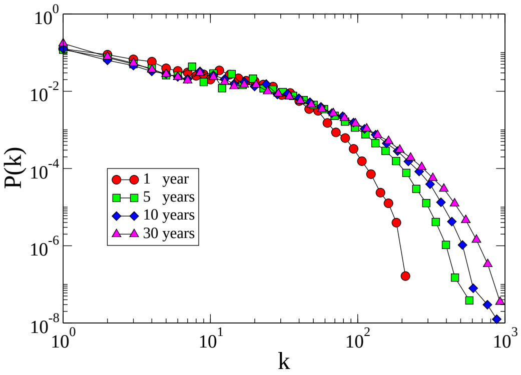

Figure S5: Considering the subset of actors active during the period 1970 − 1980 in the IMDb data we plot the degree distribution of the resulting networks for different integrating time windows.