EURECOM
Department of Mobile Communications
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report RR-13-278

Information Diffusion in Heterogeneous Networks: The Configuration Model Approach

February 13th, 2013
Last update February 13th, 2013

Pavlos Sermpezis and Thrasyvoulos Spyropoulos

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {pavlos.sermpezis, thrasyvoulos.spyropoulos}@eurecom.fr

[1]EURECOM's research is partially supported by its industrial members: BMW Group Research & Technology, IABG, Monaco Telecom, Orange, SAP, SFR, ST Microelectronics, Swisscom, Symantec.

Information Diffusion in Heterogeneous Networks: The Configuration Model Approach

Pavlos Sermpezis and Thrasyvoulos Spyropoulos

Abstract

In technological or social networks, diffusion processes (e.g. information dissemination, rumour/virus spreading) strongly depend on the structure of the network. In this paper, we focus on epidemic processes over one such class of networks, Opportunistic Networks, where mobile nodes within range can communicate with each other directly. As the node degree distribution is a salient property for process dynamics on complex networks, we use the well known Configuration Model, that captures generic degree distributions, for modeling and analysis. We also assume that information spreading between two neighboring nodes can only occur during random contact times. Using this model, we proceed to derive closed-form approximative formulas for the information spreading delay that only require the first and second moments of the node degree distribution. Despite the simplicity of our model, simulations based on both synthetic and real traces suggest a considerable accuracy for a large range of heterogeneous contact networks arising in this context, validating its usefulness for performance prediction.

Index Terms

Network Modeling, Complex Networks, heterogeneous contact dynamics, Opportunistic Networks, Configuration Model, epidemic spreading.

Contents

List of Figures

1 Introduction

Large complex networks, whether technological (e.g. Internet, World-Wide-Web, mobile P2P networks) or social (Facebook, Twitter, physical social networks), are an integral part of our lives and are becoming strongly interrelated. With information, news, videos, spam, and malware constantly spreading over such networks, it becomes increasingly interesting to understand the speed of information dissemination and its relation to network characteristics.

One such type of networks that has recently drawn a lot of attention are Opportunistic Networks [1], where mobile devices exchange data directly only when they are within wireless transmission range. As a result, messages are spread (implicitly or explicitly) in an epidemic manner, and for most applications of interest (e.g. to design and tune routing protocols, content dissemination techniques etc.), it is of main interest to estimate the spreading delay of a message.

To this end, simple epidemic models are often used, for Opportunistic Networks, so as to derive handy closed form expressions that can be used for prediction and optimization. These include, for example, simple Markovian models [2] or fluid approximations based on the celebrated SIR model used in biology and epidemiology [3]. Nevertheless, the above simple models, albeit providing useful insights, make two unrealistic assumptions: (i) that direct (regular) contacts occur between *all* pairs of nodes, and (ii) that the rate of contacts is *uniform* across all pairs. Studies of most real networks and contact traces [4, 5] reveal that neither assumption is usually true, which is consistent with our intuition (many pairs of nodes never see each other, and the rate of contacts or communication is highly heterogeneous and dependent on the mobility, social and online behavior of the agents involved).

To try to capture, in a more detailed way, the properties of real-world networks, numerous studies exist in the field of Complex Networks on epidemic processes over various complex network models (e.g. [6–10]). However, the majority of these studies focus mainly on deriving thresholds for the spreading of an epidemic disease (e.g. [6–8]) or a computer virus (e.g. [9]). Additionally, it is not always feasible to apply them in real scenarios for predicting the spreading delay, as they require the complete knowledge of the underlying contact graph [9] or the exact degree distribution [6–8]. Such information is usually very difficult, if not impossible, to estimate in real-time when considering very large networks with (possibly) time-varying topology and sparse, infrequent contacts, reducing their applicability for Opportunistic Networks. Also, some more works that consider the delay of an epidemic spreading in complex heterogeneous networks (configuration model or scale-free networks) [10, 11], have limited applicability as they derive results that can predict the message delay only for the spreading on a small, initial percentage of the total nodes of the network.

These observations leave us with the following question: *Can we still derive useful closed-form expressions that are accurate enough, even when considering more complex contact networks than usually considered for Opportunistic Net-*

works? Towards answering this question, in this paper, we remove the first assumption, namely that all nodes can potentially "infect" uniformly all other nodes. Specifically, we choose the Configuration Model [8, 12] to represent which nodes *ever* contact which others (this model can generate random instances of graphs with arbitrary degree distributions), while still assuming random contacts (with uniform rate) between nodes that do meet. Under this model, we show that we can still derive simple, closed form approximations for various quantities related to the delay of epidemic spreading (Section 2). While space limitations do not permit us to explore the second assumption analytically as well (i.e. considering different contact rates across existing links, in addition to different node degrees), we investigate its effects on the accuracy of our analytical results using simulations (Section 3). We also test our theory against real traces, capturing node mobility and respective contacts, and find that useful levels of accuracy can still be achieved even for scenarios that are known to entail considerable more complexity.

As a final remark, while our initial motivation and focus stems from the area of Opportunistic Networks, we believe that our methodology and results could also be applicable to other processes and complex networks [4], if the key metric of interest is spreading delay. In such contexts, contacts between nodes might still be subject to a random process, e.g. related to online communication (in online social networks), email transmission, etc., superimposed over a complex network (e.g. an Online Social Network friendship graph).

2 Analysis

2.1 Preliminaries

The usual way of modeling technological and social networks is through graph representation, where a link implies some sort of affinity between two nodes (e.g. online/offline friendship, actual communication link etc.). Additionally, in many situations the nodes across a given link can "contact" each other (e.g. exchange information) only at random times. For example, in Opportunistic Networks [1] nodes exchange messages only when they are within transmission range, and thus the contact times are subject to the (stochastic) mobility process of nodes. Similarly, in the case of news or videos spreading over an Online Social Network, the random contact times are dictated by the times one user would read or post something on a friends page, re-tweet, etc. [13]. To model such networks, we introduce the concept of a *Contact Network*.

Definition 1 (Contact Network)**.** *A contact network $\mathcal{N}$ is defined by (i) a (underyling) graph $\mathcal{G} = \{V, E\}$ whose vertices represent the network nodes and an edge between two vertices implies that these two nodes can contact each other regularly; (ii) each edge $i - j \in E$ is associated with a random contact process with*

rate λ_{ij}[1]. These (random) contact times, define the times during which information can be exchanged between nodes i,j.

Although the above definition is quite general, to ensure analytical tractability it is commonly assumed, either implicitly or explicitly, that the underlying graph $\mathcal{G}$ is fully meshed (i.e. a clique) or has uniform characteristics (i.e. Poisson or regular graphs) [2, 3]. Yet, in most scenarios of interest, this is rather unrealistic, as some pairs of nodes never meet and different nodes have different numbers of neighbors, resulting in a *sparse*, and largely *heterogeneous* contact graph $\mathcal{G}$.

Numerous models (e.g. [12, 14, 15]) have been proposed for representing a real network and its contact graph. One of them is the *Configuration Model* [8, 12], which creates random graphs that can have any generic degree distribution, and, thus, it can capture the degree characteristics of real-world scenarios and networks.

Definition 2 (Configuration Model)**.** *Given a network size N and a degree distribution p_d (or a degree sequence $d_i,\ i = 1, ..., N$), the Configuration Model draws random instances among all the graphs $\mathcal{G}$, with N vertices, for which the degree distribution is p_d. Connections between nodes are made randomly, and the probability of having a link between two nodes i and j is proportional only to the degrees of i and j.*

The main strengths of the Configuration Model are that: (i) It can describe networks in which the degrees[2] of the vertices can follow any arbitrary distribution. The degree distribution of the vertices[3] is an important characteristic of contact networks and it can determine the evolution of processes on the network (e.g. whether information, a virus, or a disease manages to spread.) [4]; (ii) It is based on random graphs and thus the network can be studied using analytic methods, which is the goal of our work.

Summarizing, in this paper we will consider a contact graph $\mathcal{G}$ generated by the Configuration Model, with a degree distribution p_d, and mean value and variance μ_d and σ_d^2, respectively. The coefficient of variation of the degree distribution is defined as $CV_d = \frac{\sigma_d}{\mu_d}$.

The above contact graph defines which nodes *ever* contact each other. For nodes that do contact each other, we will assume throughout our analysis that contact events are independent and identically Poisson distributed with the same rate λ for all pairs. This is a common assumption in most related analytical works [2, 3]. While the Poisson assumption might or might not be a good approximation for contact times, depending on the application scenario, it allows the use of a Markov Chain to model the spreading process as in Fig. 1. Furthemore, extensive work in the field of Opportunistic Networks [16, 17] suggest that *inter-contact* time intervals often exhibit an exponential tail. In contrast, the assumption of identical

[1]For simplicity, we will assume that the duration of each contact is quite small compared to the mean intercontact time, and thus the random contact process is a point process.

[2]The degree of a vertex is the number of edges connected to it.

[3]We will use the terms *vertex* and *node* interchangeably.

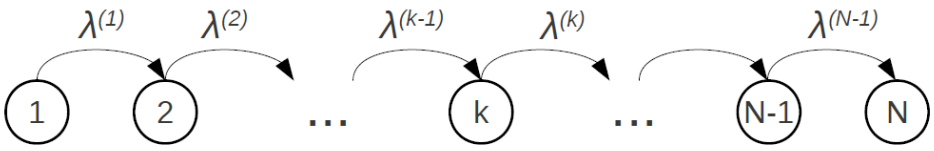


Figure 1: Epidemic spreading over a Configuration Model contact network with N nodes. The rate of moving from state k to state $k+1$ is $\lambda^{(k)}$.

contact rates λ for each node pair usually does not hold. While, we can already state that the approximation we propose is good even for (large) networks with heterogeneous contact rates (different λ_{ij} for each pair $i-j$) as well, it is beyond the scope of this paper to develop the necessary theory. Instead, we will test this approximation with simulations (Section 3).

2.2 Epidemic Spreading Model

We will now assume that a *message* is spread "epidemically" over a contact network $\mathcal{N}$, defined earlier and consisting of N total nodes. The *message* can be a data packet with useful content or a virus[4]. Specifically, we assume that a randomly chosen node x_1 generates a message and starts spreading it through the network during contacts with peers. Every node that receives the message, can further spread it to every other node that has not received it yet, when a contact with the latter occurs. To compute the expected message delivery delay of different dissemination mechanisms (e.g. routing protocols, content sharing schemes), we need to split the spreading process in steps, compute the delay of each one of these steps, and use them as the building blocks to calculate the total delay.

We say that the spreading process is at *state* k, $k = 1, ..., N-1$ when k nodes have the message, as shown in Fig. 1. We will refer to the transition from state k to state $k+1$ as *step* k. We are interested in deriving the mean delay of each such step k, starting at the time when the k^{th} node just received the message (i.e. *any* k nodes are infected) until the $(k+1)^{th}$ node receives it (i.e. *any* $k+1$ nodes are infected). We denote the set of the "infected" nodes as $\mathbf{C}(k)$. Due to the memoryless property of the Poisson contact events, the duration of step k only depends on the sum of contact rates between nodes with the message ($\in \mathbf{C}(k)$) and nodes that have not received it yet ($\notin \mathbf{C}(k)$). In Fig. 1, the sum of these rates is denoted as $\lambda^{(k)}$.

In our model, the contact rates have the same value λ for all node pairs. Hence, $\lambda^{(k)}$ is given by

$$\lambda^{(k)} = \lambda \cdot D^{out}(k) = \lambda \cdot \sum_{i \in \mathbf{C}(k)} \sum_{j \notin \mathbf{C}(k)} \mathbb{I}_{ij} \tag{1}$$

[4]In other kind of networks, it can also be a rumour or news in an Online Social Network [13], a virus in a computer network [9], a disease in the plysical world [6] etc.

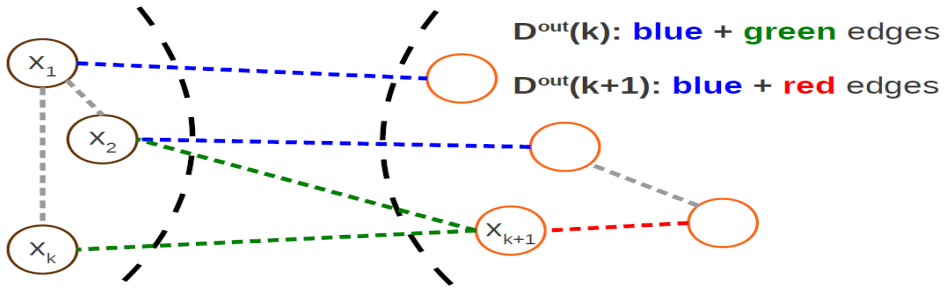


Figure 2: Sets of nodes with (left) and without (right) the message at state k. Nodes are represented by circles and edges by the straight lines.

where $\mathbb{I}_{ij} = 1$ iff there exists an edge between nodes $i - j$ (i.e. i and j contact each other). $D^{out}(k) = \sum_{i \in \mathbf{C}(k)} \sum_{j \notin \mathbf{C}(k)} \mathbb{I}_{ij}$ is defined as the *out degree* of step k. In other words, the *out degree* is the number of all the possible ways that the message can infect one additional node, when at state k.

Knowing $D^{out}(k)$ is enough to derive the total delay of each step. In a network where all the nodes contact each other (as in the usual epidemic models), it is easy to see that there are $D^{out}(k) = k(N-k)$ such $i - j$ node pairs and that the mean delay of step k in this simple case is $\frac{1}{\lambda^{(k)}} = \frac{1}{k(N-k)\lambda}$ [2]. However, in a Configuration Model network, the number of $i - j$ node pairs that could further spread the message at step k is at most $k(N-k)$, and thus the delay per step is larger. In fact, $D^{out}(k)$ is a random variable which depends on the degrees of the k nodes that *happen* to get infected first, as shown in Fig. 2. What is more, unlike uniform degree models, not all nodes here have the same probability of being infected first: nodes with higher degrees clearly have a bigger chance than nodes with low degrees. These observations complicate the derivation of step-wise delay considerably, for our more general model.

Consequently, in order to be able to derive the rate $\lambda^{(k)}$ and the mean delay of step k, it does not suffice to only know k, the number of infected nodes. We also need to keep track of the (expected) degrees that the infected nodes have at state k. Specifically, as we will show in the next sections, we need to derive the following quantities related to spreading over a configuration contact graph: (i) the expected degree of the next node to receive the message at state k, $\mu_d^{new}(k)$; and (ii) the *out degree* at step k, $D^{out}(k)$.

2.3 Mean Degree

Assume we are at state k. Let us denote as $p_d(k)$ the degree distribution of the $N - k$ nodes that *do not* have the packet at state k and $\mu_d(k)$ and $CV_d(k)$ its expectation and coefficient of variation respectively[5]. As we mentioned, not all

[5]The values of these quantities before the beginning of the spreading, are equal to the values of the initial distribution, i.e. $p_d(0) = p_d$, $\mu_d(0) = \mu_d$ and $CV_d(0) = CV_d$.

(uninfected) nodes are equally likely to be the next one infected. As a result, the expected degree of the next infected node is neither equal to μ_d (the original mean degree) nor $\mu_d(k)$.

Result 1. *The expected degree of the next node that will receive the message at step k, is approximately given by*

$$\mu_d^{new}(k) = \mu_d \cdot \left(\frac{N-k-1}{N-1}\right)^{CV_d^2} \cdot (1 + CV_d^2) \tag{2}$$

Proof. To derive the above result, we need to define and solve an appropriate recursion. Observe that there are $D^{out}(k)$ links across which the infection may proceed from state k to $k+1$ (see Fig. 2) and each of these occurs with equal probability. It is a standard result in complex network analysis [4] that the degree distribution of the node reached from that link (i.e. the next node which will receive the message) is:

$$p_d^{new}(k) = \frac{d \cdot p_d(k)}{\sum_d d \cdot p_d(k)} = \frac{d}{\mu_d(k)} \cdot p_d(k) \tag{3}$$

Eq. (3) implies that the higher degree d a node has, the more probable is that this node will be the next node to receive the message: the probability the new node to have degree d is proportional to $d \cdot p_d(k)$. Now, we can easily derive $\mu_d^{new}(k)$:

$$\mu_d^{new}(k) = \sum_d d \cdot p_d^{new}(k) = \mu_d(k) \cdot \left[1 + CV_d^2(k)\right] \tag{4}$$

We can see that the expected degree of the next node infected is higher than the mean degree of all uninfected nodes: $\mu_d^{new}(k) \geq \mu_d(k)$.

To proceed further, we thus need to know $\mu_d(k)$ and $CV_d^2(k)$ first. To this end, we can set up a recursion for the degree distribution $p_d(k)$ of the nodes that do not have the message in the next state. Notice that the set of the nodes without the message in state $k+1$ is the same set as in the previous state k, except for the node that just received the message. Hence, we can write for the number of nodes with degree d in states k and $k+1$:

$$[N - (k+1)] \cdot p_d(k+1) = (N-k) \cdot p_d(k) - p_d^{new}(k) \tag{5}$$

Substituting in Eq. (5) the value of $p_d^{new}(k)$ from Eq. (3), we find:

$$p_d(k+1) = \frac{N-k}{N-(k+1)} p_d(k) - \frac{1}{N-(k+1)} \frac{d}{\mu_d(k)} p_d(k) \tag{6}$$

In Eq. (6), we have expressed $p_d(k+1)$ as a function of $p_d(k)$. Now, it is straightforward to do the same for the expected value, $\mu_d(k+1)$, and the recursive relation for it, is:

$$\mu_d(k+1) = \mu_d(k) \cdot \left(1 - \frac{CV_d^2(k)}{N-(k+1)}\right) \tag{7}$$

To calculate $\mu_d(k+1)$, the value of $CV_d^2(k)$ is also needed. While we could also set up a recursion to derive the latter, it is proved in Appendix 5.1 that it requires knowledge of all higher moments of the degree distribution. To keep things simple and avoid requiring such knowledge (beyond the second moment), we will assume that $CV_d(k) = CV_d \ \ \forall k$. The conditions for this assumption and its accuracy are discussed in Appendix 5.1 and, here, we will only mention the main points which are: (i) the approximation can be accurate for steps k for which it holds $N - k \gg CV_d$, and (ii) it becomes more accurate as the CV_d decreases.

Thus, using $CV_d(k) = CV_d$, and $\mu_d(0) = \mu_d$, Eq. (7) gives

$$\mu_d(k) = \mu_d \cdot \prod_{m=0}^{k-1} \left(1 - \frac{CV_d^2}{N-m-1}\right) \tag{8}$$

To find an equivalent closed-form expression for Eq. (8), we can use the Taylor series approximation for the function $f(x) = e^{-x}$, about $x = 0$, which is $\mathcal{T}(e^{-x}) \approx 1 - x$ and is quite accurate for values $0 < x < 0.5$ (with increasing accuracy as x decreases). Then, setting $x = \frac{CV_d^2}{N-m-1}$ (the accuracy condition is satisfied for the states k for which $N - k > 2 \cdot CV_d^2$ and thus more accuracy can be achieved for lower values of CV_d), we can write for Eq. (8)

$$\begin{aligned}
\mu_d(k) &\approx \mu_d \cdot \prod_{m=0}^{k-1} e^{-\frac{CV_d^2}{N-m-1}} \\
&= \mu_d \cdot exp\{-CV_d^2 \cdot \sum_{m=0}^{k-1} \frac{1}{N-m-1}\} \\
&= \mu_d \cdot exp\{-CV_d^2 \cdot \sum_{\ell=N-k}^{N-1} \frac{1}{\ell}\} \\
&\approx \mu_d \cdot exp\{-CV_d^2 \cdot [\ln(N-1) - \ln(N-k-1)]\} \\
&= \mu_d \cdot exp\{\ln\left[\left(\frac{N-k-1}{N-1}\right)^{CV_d^2}\right]\} \\
&= \mu_d \cdot \left(\frac{N-k-1}{N-1}\right)^{CV_d^2}
\end{aligned} \tag{9}$$

where we have used the *harmonic series* approximation[6], which holds for $N-k \gg 1$ and whose accuracy increases for larger values of $N - k$.

Substituting Eq. (9) in Eq. (4) gives us Result 1. □

2.4 Out Degree

Result 2. *The mean value of the out degree at step k, $D^{out}(k)$, is approximately given by the relation*

[6] $\sum_{n=1}^{k} \frac{1}{n} \approx \ln(k) + \gamma$, where γ is the Euler-Mascheroni constant.

$$\overline{D}^{out}(k) = (N-k)\mu_d \left[\left(\frac{N-k}{N-1}\right)^{CV_d^2} - \left(\frac{N-2}{N-1}\right)\left(\frac{N-k}{N-1}\right)^{2CV_d^2+1} \right] \quad (10)$$

To derive Result 2 we have followed a similar method as before to form a recursion:

$$\begin{aligned} \overline{D}^{out}(k+1) &= \overline{D}^{out}(k) + [\mu_d^{new}(k) - 2] \\ &\quad -2\left[\overline{D}^{out}(k) - 1\right] \frac{\mu_d^{new}(k) - 1}{(N-k)\cdot\mu_d(k) - 1} \end{aligned} \quad (11)$$

Due to space limitations, the details about the setup and solution of Eq.(11) are omitted and can be found in Appendix 5.2. We will only provide here an intuitive sketch of proof based on a simple example.

In Fig. 2, the set of nodes with the message is $\mathbf{C}(k) = \{x_1, ..., x_k\}$ and the out degree of step k is given by the number of edges that connect the nodes $\in \mathbf{C}(k)$ with the nodes $\notin \mathbf{C}(k)$ (blue+green edges). If we denote as x_{k+1} the next node to receive the message and assume that the node x_2 disseminates the message to x_{k+1}, the out degree of the next step, $D^{out}(k+1)$, is calculated as following:

From the value of $D^{out}(k)$ we have to subtract the number of edges that connect the nodes $\in \mathbf{C}(k)$ with the node x_{k+1} (green edges). Let us denote this number as N_1. Then we have to add the number of the edges of the new node x_{k+1} that connect it with the nodes $\notin \mathbf{C}(k)$ (red edges) and we denote this number as N_2. It is evident that $N_2 = d^{new} - N_1$, where d^{new} is the degree of the node x_{k+1}. So we can write:

$$\begin{aligned} D^{out}(k+1) &= D^{out}(k) - N_1 + N_2 \\ &= D^{out}(k) + d^{new} - 2\cdot N_1 \end{aligned} \quad (12)$$

To estimate the number of the edges that connect the nodes $\in \mathbf{C}(k)$ with the node x_{k+1} (green edges), i.e. N_1, we should consider that each of the edges of $D^{out}(k)$, except for the one that connected to x_{k+1}, is connected with another edge of x_{k+1} with probability $\frac{d^{new}(k)-1}{(N-k)\cdot\mu_d(k)-1}$, where $d^{new}(k) - 1$ is the number of the unoccupied edges of x_{k+1} and $(N-k)\cdot\mu_d(k) - 1$ is the total number of edges of the nodes $\notin \mathbf{C}(k)$. We do not take into account the probability of double edges or self-loops, because this probability for large networks is almost zero [4]. So the expectation of N_1 will be

$$E[N_1] = 1 + (D^{out}(k) - 1)\cdot\frac{d^{new}(k) - 1}{(N-k)\cdot\mu_d(k) - 1} \quad (13)$$

Now, from equations Eq. (12) and Eq. (13), we can prove Eq. (11). Furthermore, using Result 1 and assuming that the minimum degree, d_{min}, of the network is

much larger than 1, which also implies that $\mu_d^{new}(k), D^{out}(k) \geq d_{min} \gg 1$, we can write for Eq. (11):

$$\overline{D}^{out}(k+1) = \overline{D}^{out}(k) \cdot \left[1 - 2\frac{1+CV_d^2}{N-k}\right] + (1+CV_d^2) \cdot \mu_d(k) \tag{14}$$

The solution of Eq. (14), for $D^{out}(1) = \mu_d$, is the Result 2.

Piecewise Formula: The above result provides us with a closed form expression for the *mean* value of the out degree $D^{out}(k)$, at step k, which allows us to calculate the necessary transition rates $\lambda^{(k)}$ in Eq.(1). However, it is based on Eq. (9) that was derived using some assumptions ($N-k \gg 1$ and $CV_d(k) = CV_d$), under which we tend to underestimate $\mu_d(k)$. Specifically, for some distributions p_d, Eq. (9) might produce, in the last steps of the recursion, unacceptably small values for $\mu_d(k)$. We can easily correct this by explicitly forcing $\mu_d(k) \geq d_{min}$ (which always holds). Then, it can be proved (Appendix 5.3) that a better approximation for $D^{out}(k)$ is given by the following piecewise result:

Result 3. *The mean value of the out degree is calculated by Eq.* (10) *for $k \leq k_{stop}$, and by*

$$\overline{D}^{out}(k) = (N-k)^2 \cdot \left[\frac{D_{stop} - d_{min} \cdot (N - k_{stop})}{(N-k_{stop})^2} + \frac{d_{min}}{N-k}\right] \tag{15}$$

for $k > k_{stop}$, where $k_{stop} = \left[1 - \left(\frac{d_{min}}{\mu_d}\right)^{\frac{1}{CV_d^2}}\right] \cdot (N-1)$ and D_{stop} is computed by setting $k = k_{stop}$ in Eq. (10).

2.5 Spreading Delay

To conclude our derivation, let us look back at our initial equation for the rates of Fig.1, $\lambda^{(k)} = \lambda \cdot D^{out}(k)$. Note that we have derived thus far the *expected* value for $D^{out}(k)$. Yet, $D^{out}(k)$ is a random variable depending on $\mathbf{C}(k)$, the actual set of the k nodes that have the message at state k. *Given* $\mathbf{C}(k)$, the delay of step k, $T_{k,k+1}$, is an exponential random variable with rate $\lambda^{(k)} = \lambda \cdot D^{out}(k)$. Thus,

$$E\left[T_{k,k+1} | \mathbf{C}(k)\right] = \frac{1}{\lambda \cdot D^{out}(k)}, \tag{16}$$

and using the properties of conditional expectation, we get the expected delay of step k:

$$E\left[T_{k,k+1}\right] = \sum_{\mathbf{C}(k)} \frac{1}{\lambda \cdot D^{out}(k)} \cdot P\{\mathbf{C}(k)\} = \frac{1}{\lambda} \cdot E\left[\frac{1}{D^{out}(k)}\right] \tag{17}$$

We cannot, in general, replace $E\left[\frac{1}{D^{out}(k)}\right]$ above, which is hard to calculate, with $\frac{1}{\overline{D}^{out}(k)}$, which follows directly from Eq.(10) and (15). In fact, Jensen's inequality suggests that $\frac{1}{\overline{D}^{out}(k)} \leq E\left[\frac{1}{D^{out}(k)}\right]$.

To proceed with our approximation, we resort to the *Delta method* [18]. This is a method for approximating the expectation of functions of random variables. Here, the random variable is $X = D^{out}(k)$ and we need to compute (Eq. (17)) the expectation of the function $f(X) = \frac{1}{X}$. We can approximate $f(X)$ with a Taylor series expansion about the mean value $E[X] = \overline{D}^{out}(k)$. Finally, by keeping only the first few terms of this series and taking their expectation, we can more easily express $E[T_{k,k+1}]$ as a function of moments of $D^{out}(k)$. Specifically, considering the first two terms of the expansion, we get

$$E\left[T_{k,k+1}\right] = \frac{1}{\lambda} \cdot E\left[\frac{1}{D^{out}(k)}\right] \approx \frac{1}{\lambda \cdot \overline{D}^{out}(k)} \tag{18}$$

Now, in Eq. (18), we can calculate the expected *step delay* by substituting the value of Eq. (10) or Eq. (15).

The accuracy of the Delta method and the above approximation is higher, if the mass of the random variable $X = D^{out}(k)$ is concentrated around its mean $\overline{D}^{out}(k)$ [18]. It is known that, in a configuration model network, the network structural properties and the properties of processes on the network becoming concentrated more and more narrowly around their mean value [4], as the network size increases. Therefore, the larger the network size N, the higher the accuracy of the approximation. Furthermore, if increased accuracy is desired, more terms in the Taylor series above could be used (by deriving a few higher moments of $D^{out}(k)$).

3 Model Validation

In order to validate our model, we compare the theoretical results we derived, against a sample of simulations for both synthetic and real-world networks.

Synthetic Simulations: At first, we created various synthetic scenarios conforming to our model (Section 2.1). For each scenario, the procedure we follow, is:

1. We choose an initial degree distribution p_d.
2. With the configuration model we create 50 different networks (contact graphs) and for each pair of nodes in a network we create a sequence of contact events with inter-contact times drawn from an exponential distribution with rate $\lambda = 1$.
3. For each network, we generate 1000 messages at random times and at random source nodes and start the speading.

4. We calculate the average values, over all networks and spreading processes of the specific scenario, of the *out degree*, $\overline{D}^{out}(k)$, and *step delay*, $E[T_{k,k+1}]$, of each step.

To choose realistic parameters for the degree distributions in our scenarios, we analysed contact graphs of real-world networks[7] and found that the degrees follow either a uniform or right-skewed distribution with CV_d in the range $[0.6, 0.85]$ (details for the scenarios in Table 3).

Table 1: Parameters of the contact graphs of four real-world scenarios.

TRACE	network size N	μ_d	CV_d
Sigcomm 2009	76	25.5	0.6
SocioPatterns	111	7.6	0.85
Cabspotting	536	120	0.74
Infocom 2006	98	32	0.61

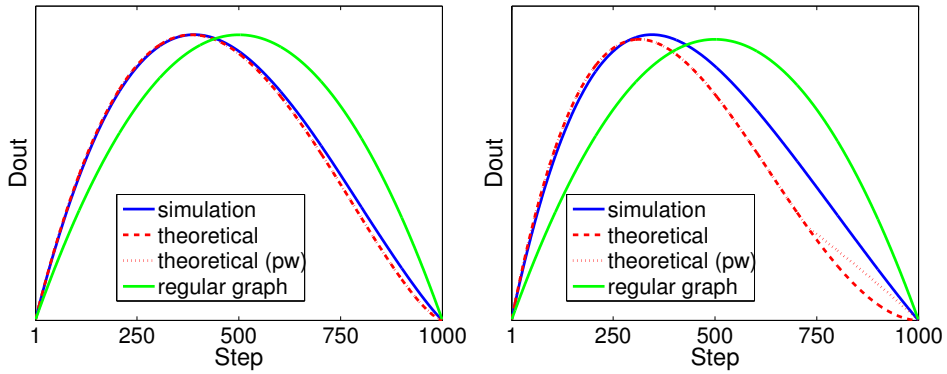
In Fig. 3 we present the *out degree* for each step in two scenarios with 1000 nodes. We compare the simulation values with the theoretical (Results 2 and 3). We can see that the achieved accuracy is significant. As expected, in the scenario with higher CV_d the accuracy is lower, especially for the last steps, because the approximations we did in the derivation of the theoretical results are less accurate as the CV_d increases. Also, in Fig. 8(a) there was not need to use the piecewise formula (Result 3) and in the second case, Fig. 8(b), it should be used only for the last 25% of the steps. The corresponding values for $D^{out}(k)$ that a *fully-meshed* network model would predict are very far from the simulated values (e.g. for the 500^{th} step it gives a value 15 times larger). We therefore also compare our results to a baseline model: a *regular graph* with the same number of edges as our network, but where every node has the same degree. Fig. 3 confirms that our model performs significantly better.

In Table 2 we present the *average relative errors* for $D^{out}(k)$, defined as

$$E\left[\frac{|D^{out}(k)_{sim} - D^{out}(k)_{th}|}{D^{out}(k)_{sim}}\right]$$

, for four networks (of which the two correspond to the results presented in Fig. 3) of 1000 nodes and similar μ_d values. We show the average relative error for the first 250, 500 and 750 steps and the total (over all steps). The more steps we consider, the higher the error is. This comes of the fact that our theoretical results are less accurate for the last steps of the spreading (Section 2). It can be seen that

[7]The traces are available at:
1) *SocioPatterns*: http://www.sociopatterns.org/
2) *Sigcomm 2009*: http://crawdad.cs.dartmouth.edu/thlab/sigcomm2009
3) *Cabspotting*: http://crawdad.cs.dartmouth.edu/epfl/mobility
4) *Infocom 2006*: http://crawdad.cs.dartmouth.edu/cambridge/haggle



(a) $CV_d = 0.65$ (b) $CV_d = 0.92$

Figure 3: $D^{out}(k)$ of each step in two scenarios with 1000 nodes.

for networks with lower CV_d the error is lower. For example, for $CV_d = 0.31$, the error is insignificant, even for the last steps. For the extreme case of $CV_d = 1.29$ [8] the error is not negligible. However, our prediction is still acceptable, if we consider the heterogeneity this scenario has.

Table 2: Relative step error of $D^{out}(k)$ on different network scenarios.

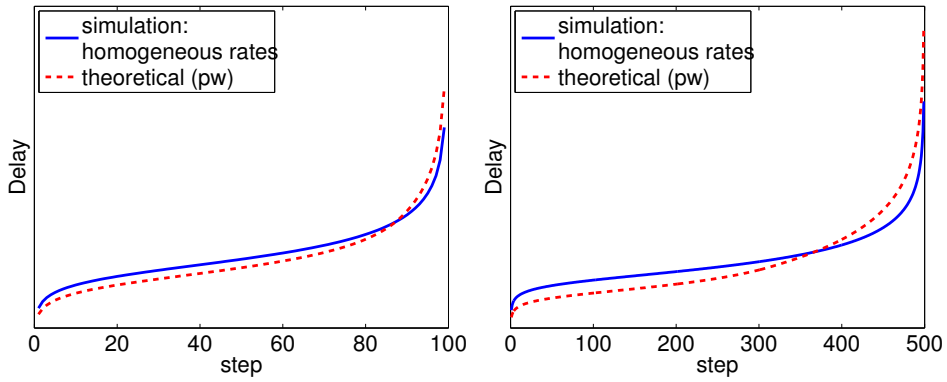
	250 steps	500 steps	750 steps	over all steps
$CV_d = 0.31$	1%	2%	2%	2%
$CV_d = 0.65$	1%	1%	2%	6%
$CV_d = 0.92$	4%	4%	11%	15%
$CV_d = 1.29$	14%	18%	27%	29%

Fig. 4 shows the *aggregate step delay* (i.e. the time the message needs to be spread in k nodes) for two synthetic scenarios: (a) network with with 100 nodes, $\mu_d = 23$ and $CV_d = 0.71$; and (b) network with 500 nodes, $\mu_d = 30$ and $CV_d = 1.16$ [9]. Similarly to the results for $D^{out}(k)$, it can be seen also here that the theoretical *aggregate step delay* is close to the simulated value for almost every step.

Synthetic Simulations - Heterogeneous Rates: Further, we investigate the performance of our model in networks with heterogeneous contact rates (different λ_{ij} for each pair). We create synthetic scenarios and run simulations as before. The only difference is the generation of the contact events, where, now, the inter-contact times are exponentially distributed but with a different rate for each pair. We chose λ_{ij} to follow a log-normal distribution with $\mu_\lambda = 1$ and $\sigma_\lambda^2 = 3$.

[8] We characterise it as an extreme case, as the min and max degrees in this network are 22 and 968, respectively, in order to have a CV_d value as high as possible.

[9] It is the higher variance we could achieve among all the scenarios of 100 and 500 nodes, respectively. The degree distribution was highly skewed and the maximum degree in the network was almost equal to the network size, $d_{max} = 100$ and $d_{max} = 500$ for the two cases.



(a) $\mu_d = 23$, $CV_d = 0.71$ (b) $\mu_d = 30$, $CV_d = 1.16$

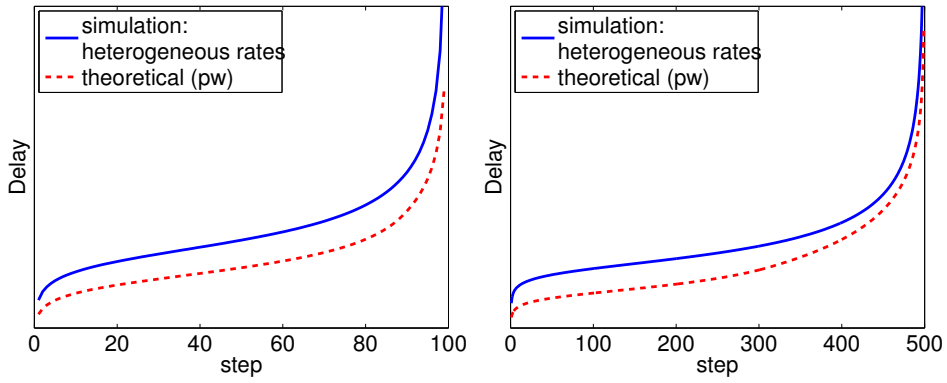
Figure 4: Aggregate step delay. Synthetic simulations in scenarios with: (a) 100 nodes, $\mu_d = 23$ and $CV_d = 0.71$; and (b) network with 500 nodes, $\mu_d = 30$ and $CV_d = 1.16$.

The results for the *aggregate step delay* are presented in Fig. 5. The scenarios presented are the corresponding to the homogeneous-rates scenarios of Fiq. 4. As can be seen in Fig. 5, simulation and theoretical results diverge more for the heterogeneous contact rate scenario.

This divergence is more clearly seen in Fig. 6, which shows the relative error of the average *aggregate step delay* over all the steps, i.e. $E\left[\frac{|D_{sim}-D_{th}|}{D_{sim}}\right]$ where D denotes the aggregate step delay. We present six scenarios of different network sizes. For each scenario we chose a bounded pareto degree distribution with minimum value $d_{min} = 0.1 \cdot N$ (N is the network size), $d_{max} = N$ and shape factor the one that resulted in the higher CV_d. These represent the worst case parameters (among the ones we observed in real traces) that most hurt the accuracy of our model. Nevertheless, in the homogeneous scenarios, the error is very low (below 10% for almost all the networks) and, in the heterogeneous scenarios, it is always higher, but decreases for larger network sizes. For a network with 300 nodes, it becomes approximately 20%, which is rather satisfying, given the high variability in both the degrees and rates in this scenarios.

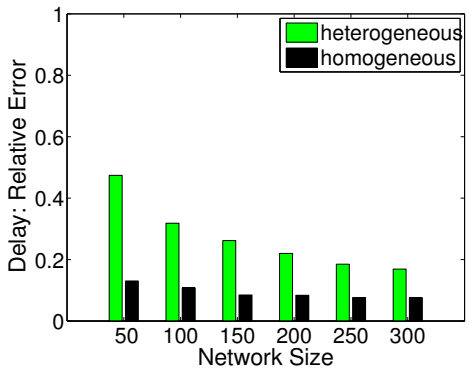
Real-world Networks: After evaluating the accuracy of our model in a range of different (regarding the network size, degree distribution, contact rates) synthetic scenarios, we present here the results of simulations on real-world traces. It is of interest to see to what extent our model can capture the quantities of interest in a real-world scenario, where the assumptions do not hold exactly, as we have noted community structure (i.e. the *clustering coefficient* [4] is $27 - 50\%$ more than in the corresponding configuration model network), heterogeneous contact rates and non-Poisson contact events (e.g. less contacts during night hours).

Fig. 7 shows the results of 1000 simulation runs on the mobility trace from the 4 days iMotes experiment during Infocom 2006 [19], which contains traces of



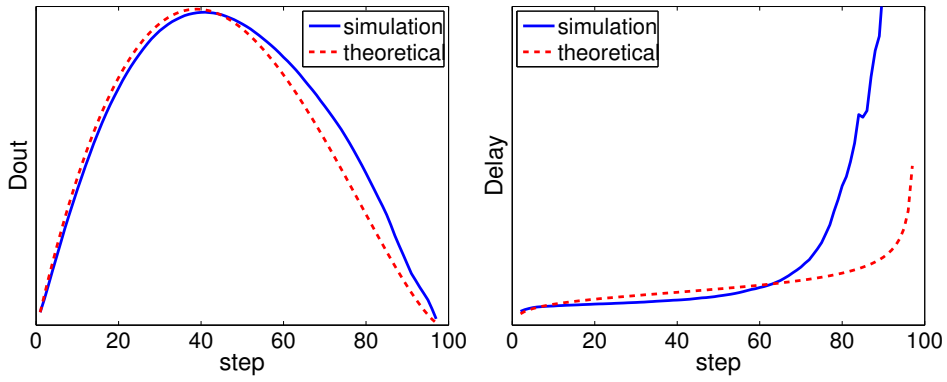
(a) $\mu_d = 23$, $CV_d = 0.71$ (b) $\mu_d = 30$, $CV_d = 1.16$

Figure 5: Aggregate step delay. Synthetic simulations in scenarios with heterogeneous contact rates: (a) 100 nodes, $\mu_d = 23$ and $CV_d = 0.71$; and (b) network with 500 nodes, $\mu_d = 30$ and $CV_d = 1.16$.



(a) Delay: Relative Errors

Figure 6: Relative errors of the delay averaged over all the steps in scenarios with Homogeneous and Heterogeneous contact rates for 6 different network sizes.



(a) $D^{out}(k)$ (b) Delay

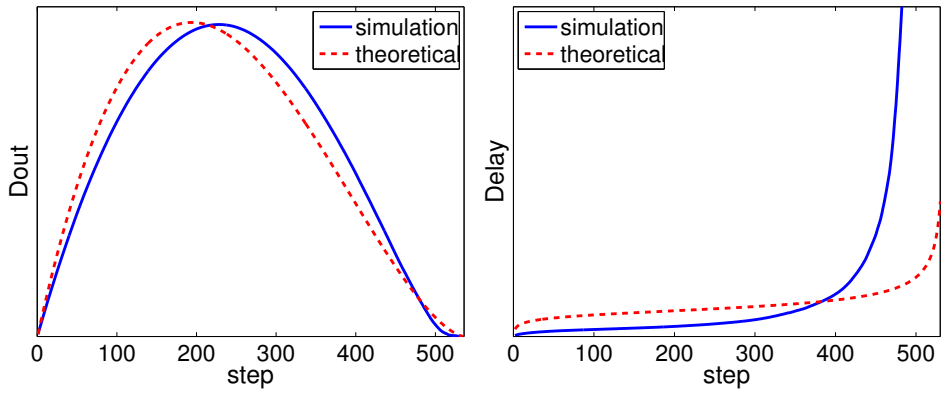
Figure 7: Simulations on Infocom 2006 trace: 96 nodes, $\mu_d = 33$, $CV_d = 0.6$

Bluetooth sightings of 78 mobile and 20 static nodes. In Fig. 7(a) it can be seen that the theoretically predicted *out degree* only differ slightly, except for some last steps, from the simulation's average. Thus we can infer, that despite the community structure of this network, our model can still capture the way the spreading proceeds among nodes with different degree. Fig. 7(b) shows the *aggregate step delay*. We can see that the accuracy is good for more than half of the steps. However, in the following steps our theoretical results are far from the observed delay. An explanation for this, is the correlation between the contact events of different pairs which affects the spreading process (e.g. in conference events there are much more contact events than during night hours).

We have observed similar good accuracy for the first 70-75% steps and divergence subsequently, in other traces as well. In Fig. 8 we present the results of 1000 simulation runs on the mobility trace *Cabspotting* [20], which contains GPS coordinates from 536 taxi cabs collected over 30 days in San Francisco.

4 Conclusions

In this paper, we have derived closed form approximations for the step-wise and total delay of epidemic spreading on graphs with arbitrary degree distributions, where neighbors contact randomly. Despite the assumptions made, and the use of only partial knowledge of the degree distribution, we conclude that our results offer useful accuracy in networks with reasonable heterogeneity (CV_d), which is the case for many Opportunistic Networks. Even for some real contact network examples, known to exhibit considerably more structure, our result provides very good accuracy for over 70% of the spreading process. However, some social networks exhibit much higher values of heterogeneity (CV_d). This means that, for such networks, such closed form approximations might not be feasible. Coarser bounds (e.g. based on conductance and spectral analysis) might offer an alterna-



(a) $D^{out}(k)$ (b) Delay

Figure 8: Simulations on Cabspotting trace: 536 nodes, $\mu_d = 120$, $CV_d = 0.74$

tive, but they come at the cost of potentially large errors and prohibitive complexity for online and distributed estimation, especially in a low contact rate context. We intend to consider such tradeoffs further as a part of future work.

5 Appendix

5.1 Conditions for Assuming a Constant Coefficient of Variation

From Eq. (6), we can easily result to the recurrence relation for the second moment of the degree distribution:

$$\overline{d^2}(k+1) = \frac{N-k}{N-(k+1)}\overline{d^2}(k) - \frac{1}{N-(k+1)}\frac{\overline{d^3}(k)}{\mu_d(k)} \tag{19}$$

where $\overline{d^n}(k)$ is the n^{th} moment of the degree distribution.

As we have computed the expectation and the second moment of the degree distribution in step $k+1$, Eq. (7) and Eq. (19) respectively, we can find the recurrence relation for the coefficient of variation, which is:

$$CV_d^2(k+1) = \frac{CV_d^2(k)\cdot\left(1-\frac{\gamma_d(k)\cdot CV_d(k)+2}{N-k-1}\right)+1}{\left(1-\frac{CV_d^2(k)}{N-k-1}\right)^2} - 1 \tag{20}$$

where we denote as $\gamma_d(k)$ the skewness of the degree distribution. In Eq. (20), if we do not know the value of$\gamma_d(k)$, we cannot solve the recurrence relation for $CV_d(k)$ and we cannot evaluate it. Thus, as we can see, the expression for the value of $CV_d^2(k)$ (which is equivalent to the second moment $E[d^2(k)]$) includes the value of the third moment of the degree distribution at state k. So, recursively, it follows that the exact solution of Eq. (7) requires the knowledge of all the higher moments of

the degree distribution. However, this requirement both increases complexity and decreases applicability as it is not always efficient or possible to know or estimate all the higher moments of the degree distribution. Therefore, in order to find a closed form solution for $\mu_d(k)$, we assume $CV_d(k) = CV_d \ \forall k$.

This relation hold for the cases where $\frac{\gamma_d(k) \cdot CV_d(k)+2}{N-k-1} \ll 1$ and $\frac{CV_d^2(k)}{N-k-1} \ll 1$, where it is easy to see from Eq. 20 that

$$CV_d^2(k+1) \simeq CV_d^2(k) \tag{21}$$

Summarizing, it is relatively accurate to assume that the coefficient of variation of the degree distribution remains the same for each state k, when

$$N - k \gg max\{1, \ CV_d^2, \ \gamma_d \cdot CV_d^2\} \tag{22}$$

5.2 Proof of Result 2

5.2.1 Rigorous Proof of the recurrence relation, Eq. (11)

At first we will show, rigorously, how we derived the recurrence relation for the mean *out degree* in each step, i.e. $\overline{D}^{out}(k)$.

Proof. At step k, the average degree of the nodes that do not have the message is $\mu_d(k)$ and is given by Eq. (9). Thus, it holds that the total number of edges, which are not connected to $\mathbf{C}(k)$ is $(N-k) \cdot \mu_d(k)$. Let the *out degree* to be $D^{out}(k)$ and the degree of the next node to receive the message to be $d^{new}(k)$ [10]. According to the reasoning of Section 2.4, the *out degree* of the next step will be

$$D^{out}(k+1) = D^{out}(k) + (d^{new}(k) - 2) - 2 \cdot \mathcal{H}(M, m, n) \tag{23}$$

where $\mathcal{H}(M, m, n)$ is a random variable drawn from a *Hypergeometric* distribution [11] with parameters

$$\begin{aligned} M &= (N-k) \cdot \mu_d(k) - 1 \\ m &= d^{new}(k) - 1 \\ n &= D^{out}(k) - 1 \end{aligned}$$

Taking the expectation of both sides of Eq. (24) we get

$$\overline{D}^{out}(k+1) = \overline{D}^{out}(k) + (\mu_d^{new}(k) - 2) - 2 \cdot E[\mathcal{H}(M, m, n)] \tag{24}$$

The value of $\mu_d^{new}(k)$ is given by Eq. (2). We cannot calculate directly the expectation of the Hypergeometric distribution, because its arguments are random

[10] $D^{out}(k)$ and $d^{new}(k)$, are not expectations as in Section 2, but they are random variables.

[11] The *Hypergeometric* distribution is a discrete probability distribution that describes the probability of l successes in n draws from a finite population of size M, containing m successes, *without* replacement.

variables too. Therefore, we need to compute first the conditional expectation, conditioning on $D^{out}(k)$ and $d^{new}(k)$:

$$\begin{aligned}
&E\left[\mathcal{H}\left(M,m,n\right)\right] = \\
&= \sum_{D^{out'}} \sum_{d^{new'}} E\left[\mathcal{H}\left(M,m,n\right)|\, D^{out'}, d^{new'}\right] \cdot P\left(D^{out'}, d^{new'}\right) \\
&= \sum_{D^{out'}} \sum_{d^{new'}} \frac{n \cdot m}{M} \cdot P\left(D^{out'}, d^{new'}\right) \\
&= \sum_{D^{out'}} \sum_{d^{new'}} \frac{(d^{new'}-1)\cdot(D^{out'}-1)}{(N-k)\cdot\mu_d(k)-1} \cdot P\left(D^{out'}, d^{new'}\right) \qquad (25)
\end{aligned}$$

and as $D^{out}(k)$ and $d^{new}(k)$ are independent random variables, then Eq. (25) becomes

$$E\left[\mathcal{H}\left(M,m,n\right)\right] = \frac{(\mu_d^{new}(k)-1)\cdot(D^{out}(k)-1)}{(N-k)\cdot\mu_d(k)-1} \qquad (26)$$

and Eq. (24) turns into Eq. (11). □

5.2.2 Solution of Eq. (14)

Now, that we derived the recurrence relation Eq. (11), we solve its equivalent expression which is given by Eq. (14).

Proof. For $k = 1$, Eq. (14) gives:

$$\overline{D}^{out}(2) = \overline{D}^{out}(1) \cdot \left[1 - 2\frac{1+CV_d^2}{N-1}\right] + (1+CV_d^2)\cdot\mu_d(1),$$

for $k = 2$, it gives:

$$\begin{aligned}
\overline{D}^{out}(3) &= \overline{D}^{out}(2) \cdot \left[1 - 2\frac{1+CV_d^2}{N-2}\right] + (1+CV_d^2)\cdot\mu_d(2) \\
&= \overline{D}^{out}(1) \cdot \left[1 - 2\frac{1+CV_d^2}{N-1}\right] \cdot \left[1 - 2\frac{1+CV_d^2}{N-2}\right] \\
&+ (1+CV_d^2)\cdot\mu_d(1) \cdot \left[1 - 2\frac{1+CV_d^2}{N-2}\right] \\
&+ (1+CV_d^2)\cdot\mu_d(2)
\end{aligned}$$

and recursively, it can be expressed as

$$\begin{aligned}
\overline{D}^{out}(k) =& \overline{D}^{out}(1) \cdot \prod_{m=1}^{k-1}\left[1 - 2\frac{1+CV_d^2}{N-m}\right] \\
&+ \sum_{m=1}^{k-1}(1+CV_d^2)\cdot\mu_d(k)\prod_{\ell=m+1}^{k-1}\left[1 - 2\frac{1+CV_d^2}{N-\ell}\right] \qquad (27)
\end{aligned}$$

To find a closed-form expression (without sums and products of many terms) for Eq. (27) we need first to calculate the sums and products separately. So, at first:

$$\begin{aligned}
&\prod_{m=1}^{k-1}\left[1-2\frac{1+CV_d^2}{N-m}\right] \\
&\quad\approx \prod_{m=1}^{k-1} e^{-2\frac{1+CV_d^2}{N-m}} \\
&\quad= exp\{-2\left(1+CV_d^2\right)\cdot\sum_{m=1}^{k-1}\frac{1}{N-m}\} \\
&\quad= exp\{-2\left(1+CV_d^2\right)\cdot\sum_{m=N-k+1}^{N-1}\frac{1}{m}\} \\
&\quad\approx exp\{-2\left(1+CV_d^2\right)\cdot[\ln(N-1)-\ln(N-k)]\} \\
&\quad= exp\{-2\left(1+CV_d^2\right)\cdot\ln\left(\frac{N-1}{N-k}\right)\} \\
&\quad= \left(\frac{N-k}{N-1}\right)^{2\left(1+CV_d^2\right)} \\
&\quad= \left(\frac{N-k}{N-1}\right)\cdot\left(\frac{N-k}{N-1}\right)^{1+2CV_d^2}
\end{aligned} \tag{28}$$

where for the first approximation we used the Taylor series expansion (similarly to the proof of Result 1), which is accurate for $N-k > 4(1+CV_d^2)$, and for the second approximation we used the harmonic series approximation, whose accuracy increases for larger values of $N-k$.

Similarly to Eq. (28), we can find that

$$\prod_{\ell=m+1}^{k-1}\left[1-2\frac{1+CV_d^2}{N-\ell}\right]\approx\left(\frac{N-k}{N-m-1}\right)^{2\left(1+CV_d^2\right)} \tag{29}$$

and now, using Eq. (29), we can write for the summation in Eq. (27)

$$\sum_{m=1}^{k-1}(1+CV_d^2)\cdot\mu_d(k)\prod_{\ell=m+1}^{k-1}\left[1-2\frac{1+CV_d^2}{N-\ell}\right]$$
$$=\sum_{m=1}^{k-1}(1+CV_d^2)\cdot\mu_d(k)\cdot\left(\frac{N-k}{N-m-1}\right)^{2\left(1+CV_d^2\right)}$$
$$=(1+CV_d^2)\sum_{m=1}^{k-1}\mu_d\left(\frac{N-m-1}{N-k}\right)^{CV_d^2}\cdot\left(\frac{N-k}{N-m-1}\right)^{2\left(1+CV_d^2\right)}$$
$$=(1+CV_d^2)\cdot\mu_d\cdot\frac{(N-k)^{2\left(1+CV_d^2\right)}}{(N-1)^{CV_d^2}}\cdot\sum_{m=1}^{k-1}\left(\frac{1}{N-m-1}\right)^{2+CV_d^2}$$
$$=(1+CV_d^2)\cdot\mu_d\cdot\frac{(N-k)^{2\left(1+CV_d^2\right)}}{(N-1)^{CV_d^2}}\cdot\sum_{m=N-k}^{N-2}\frac{1}{m^{2+CV_d^2}} \tag{30}$$

We approximate the sum that appears in the right side of the last line in Eq. (30) with the integral

$$\sum_{m=N-k}^{N-2}\frac{1}{m^{2+CV_d^2}}\approx\int_{N-k}^{N-1}\frac{1}{m^{2+CV_d^2}}dm$$
$$=\frac{(N-1)^{\left(1-(2+CV_d^2)\right)}-(N-k)^{\left(1-(2+CV_d^2)\right)}}{1-(2+CV_d^2)}$$
$$=\frac{1}{1+CV_d^2}\left[\frac{1}{(N-k)^{1+CV_d^2}}-\frac{1}{(N-1)^{1+CV_d^2}}\right] \tag{31}$$

and finally, combining Eq. (30) and Eq. (31), we get

$$\sum_{m=1}^{k-1}(1+CV_d^2)\cdot\mu_d(k)\prod_{\ell=m+1}^{k-1}\left[1-2\frac{1+CV_d^2}{N-\ell}\right]$$
$$=\mu_d\cdot\frac{(N-k)^{2\left(1+CV_d^2\right)}}{(N-1)^{CV_d^2}}\cdot\left[\frac{1}{(N-k)^{1+CV_d^2}}-\frac{1}{(N-1)^{1+CV_d^2}}\right]$$
$$=\mu_d\cdot(N-k)\cdot\left[\left(\frac{N-k}{N-1}\right)^{CV_d^2}-\left(\frac{N-k}{N-1}\right)^{1+2CV_d^2}\right] \tag{32}$$

Substituting in Eq.(27) the expressions from Eq.(28) and Eq.(32) and having in mind that $\overline{D}^{out}(1)=\mu_d$, we can write

$$\overline{D}^{out}(k)$$
$$=\mu_d\cdot\left(\frac{N-k}{N-1}\right)\cdot\left(\frac{N-k}{N-1}\right)^{1+2CV_d^2}$$
$$+\mu_d\cdot(N-k)\cdot\left[\left(\frac{N-k}{N-1}\right)^{CV_d^2}-\left(\frac{N-k}{N-1}\right)^{1+2CV_d^2}\right]$$
$$=\mu_d\cdot(N-k)\left[\left(\frac{N-k}{N-1}\right)^{CV_d^2}-\left(1-\frac{1}{N-1}\right)\left(\frac{N-k}{N-1}\right)^{1+2CV_d^2}\right] \tag{33}$$

which is the Result 2. □

5.3 Proof of Result 3

Applying the condition $\mu_d(k) \geq d_{min}$ in Eq. (9), we can find that it is satisfied for the steps k that

$$k \leq \left[1 - \left(\frac{d_{min}}{\mu_d}\right)^{\frac{1}{CV_d^2}}\right] \cdot (N-1) = k_{stop} \tag{34}$$

The previous equation means that after the k_{stop}^{th} state [12], Eq. (9), gives values $\mu_d(k) \leq d_{min}$. To overcome this problem, we will use Eq. (10) for calculating $\overline{D}^{out}(k)$ for $k \leq k_{stop}$ till step k_{stop} and then, as all the remaining nodes must have degree d_{min}, use the recurrence relation:

$$\overline{D}^{out}(k+1) = \overline{D}^{out}(k) + d_{min} - \frac{2 \cdot \overline{D}^{out}(k)}{N-k} \tag{35}$$

Solving, similarly as in Appendix 5.2, the Eq. (35), for initial condition $\overline{D}^{out}(k_{stop}) = D_{stop}$ where the value of D_{stop} is taken from Eq. (10), we end up to the recurrence relation

$$\begin{aligned}\overline{D}^{out}(k) =& D_{stop} \cdot \prod_{m=k_{stop}}^{k-1}\left(1 - \frac{2}{N-m}\right) \\ &+ d_{min} \cdot \sum_{m=k_{stop}}^{k-1} \prod_{\ell=m+1}^{k-1}\left(1 - \frac{2}{N-m}\right)\end{aligned} \tag{36}$$

for $k > k_{stop}$. Using the Teylor series expansion and Harmonic series approximations we can show that

$$\prod_{m=k_{stop}}^{k-1}\left(1 - \frac{2}{N-m}\right) = \left(\frac{N-k}{N-k_{stop}}\right)^2 \tag{37}$$

$$\prod_{\ell=m+1}^{k-1}\left(1 - \frac{2}{N-m}\right) = \left(\frac{N-k}{N-m-1}\right)^2 \tag{38}$$

[12] In case $k_{stop} > N-1$ the following analysis is not needed and we can use the Result 2

and then, by Eq. (38):

$$\begin{aligned}
&\sum_{m=k_{stop}}^{k-1} \prod_{\ell=m+1}^{k-1} \left(1 - \frac{2}{N-m}\right) \\
&\quad = \sum_{m=k_{stop}}^{k-1} \left(\frac{N-k}{N-m-1}\right)^2 \\
&\quad = (N-k)^2 \cdot \sum_{m=N-k}^{N-k_{stop}-1} \frac{1}{m^2} \\
&\quad \approx (N-k)^2 \cdot \int_{m=N-k}^{N-k_{stop}} \frac{1}{m^2} dm \\
&\quad = (N-k)^2 \cdot \left[\frac{1}{N-k} - \frac{1}{N-k_{stop}}\right] \qquad (39)
\end{aligned}$$

Now, Result 3 follows easily by substituting the expressions of Eq. (37) and Eq. (39) in Eq. (35).

Remark: As we saw, in our analysis, we first consider Result 1 and for the last steps we assume $\mu_d^{new}(k) = d_{min}$ in order to derive Result 3. In addition to the intuitive reasons, which we described, this assumption can also justified by a similar work. In [10], the authors investigate, through analysis and simulations, the average degree of the newly infected nodes, $\mu_d^{new}(k)$. They conclude that in early steps $\mu_d^{new}(k)$ is given by $\frac{\overline{d^2}}{\overline{d}} = \mu_d \cdot (1 + CV_d^2)$, which is in agreement with our result, and then it gradually decreases and in the last steps it becomes equal to the minimun degree of the network, d_{min}.

References

[1] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: data forwarding in disconnected mobile ad hoc networks," *Comm. Mag., IEEE*, vol. 44, no. 11, pp. 134–141, Nov. 2006.

[2] R. Groenevelt, P. Nain, and G. Koole, "The message delay in mobile ad hoc networks," *Performance Evaluation*, vol. 62, pp. 210–228, 2005.

[3] X. Zhang, G. Neglia, J. Kurose, and D. Towsley, "Performance modeling of epidemic routing," *Computer Networks*, vol. 51, no. 10, pp. 2867–2891, 2007.

[4] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[5] A. Passarella, M. Conti, C. Boldrini, and R. I. Dunbar, "Modelling inter-contact times in social pervasive networks," in *Proc. ACM MSWiM'11*.

[6] M. E. J. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, Jul 2002. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.66.016128

[7] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, "Epidemic outbreaks in complex heterogeneous networks," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 26, pp. 521–529, 2002. [Online]. Available: http://dx.doi.org/10.1140/epjb/e20020122

[8] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical Review E*, vol. 64, no. 2, Aug. 2001.

[9] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic spreading in real networks: An eigenvalue viewpoint." in *SRDS*. IEEE Computer Society, 2003, pp. 25–34.

[10] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, "Dynamical patterns of epidemic outbreaks in complex heterogeneous networks," *Journal of Theor. Biology*, vol. 235, no. 2, pp. 275–288, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022519305000251

[11] T. Zhou, J.-G. Liu, W.-J. Bai, G. Chen, and B.-H. Wang, "Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity," *Phys. Rev. E*, vol. 74, Nov 2006. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.74.056109

[12] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random Structures & Algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995. [Online]. Available: http://dx.doi.org/10.1002/rsa.3240060204

[13] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," in *In SDM*, 2007.

[14] P. Erdős and A. Rényi, "On random graphs. I," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.

[15] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of Sociology*, vol. 105, pp. 493–527, 1999.

[16] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnović, "Power law and exponential decay of inter contact times between mobile devices," in *Proc. ACM MobiCom*, 2007.

[17] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *Proc. of ACM MobiHoc*, 2009. [Online]. Available: http://doi.acm.org/10.1145/1530748.1530790

[18] G. W. Oehlert, "A note on the delta method," *The American Statistician*, vol. 46, no. 1, pp. 27–29, 1992.

[19] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, "CRAWDAD data set cambridge/haggle (v. 2009-05-29)," Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle, May 2009.

[20] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD data set epfl/mobility (v. 2009-02-24)," Downloaded from http://crawdad.cs.dartmouth.edu/epfl/mobility, Feb. 2009.