Analysis of the Propagation Time of a Rumour in Large-scale Distributed Systems

Yves Mocquard, Bruno Sericola, Samantha Robert, Emmanuelle Anceaume

Analysis of the Propagation Time of a Rumour in Large-scale Distributed Systems

Yves Mocquard
Université Rennes 1/IRISA,
yves.mocquard@irisa.fr

Samantha Robert
Université de Nantes,
samantha.robert@irisa.fr

Bruno Sericola
INRIA Rennes - Bretagne Atlantique,
bruno.sericola@inria.fr

Emmanuelle Anceaume
CNRS/IRISA,
emmanuelle.anceaume@irisa.fr

***Abstract*—The context of this work is the well studied dissemination of information in large scale distributed networks through pairwise interactions. This problem, originally called *rumor mongering*, and then *rumor spreading* has mainly been investigated in the synchronous model. This model relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In this paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where at time unit, a single node interacts with a randomly chosen neighbor. We perform a thorough study of T_n the total number of interactions needed for all the n nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow for comparing different protocols, we prove that in a complete graph of size $n \geq 2$, the probability that $T_n > k$ for all $k \geq 1$ is less than $\left(1 + \frac{2k(n-2)^2}{n}\right)\left(1 - \frac{2}{n}\right)^{(k-1)}$. We also study the behavior of the complementary distribution of T_n at point $c\mathbb{E}(T_n)$ when n tends to infinity for $c \neq 1$. We end our analysis by conjecturing that when n tends to infinity, $T_n > \mathbb{E}(T_n)$ with probability close to 0.4484.**

***Keywords*—rumor spreading, pairwise interactions, Markov chain, analytical performance evaluation.**

This paper should be considered for the best student paper award. Yves Mocquard is PhD student at the University of Rennes 1. E-mail: yves.mocquard@irisa.fr

I. Introduction

Randomized rumor spreading is an important mechanism that allows the dissemination of information in large and complex networks through pairwise interactions. This mechanism initially proposed by Deemers et al [11] for the update of a database replicated at different sites, has then been adopted in many applications ranging from resource discovery [17], data-aggregation [21], complex distributed applications [7], or virus propagation in computer networks [5], to mention just a few.

A lot of attention has been devoted to the design and study of randomized rumor spreading algorithms. Initially, some rumor is placed on one of the vertices of a given network, and this rumor is propagated to all the vertices of the network through pairwise interactions between vertices. One of the important questions of these protocols is the *spreading time*, that is time it needs for the rumor to be known by all the vertices of the network.

Several models have been considered to answer this question. The most studied one is the synchronous push/pull model, also called the synchronous random phone call model. This model assumes that all the vertices of the network act in synchrony, which allows the algorithms designed in this model to divide time in synchronized rounds. During each synchronized round, each vertex i of the network selects at random one of its neighbor j and either sends to j the rumor if i knows it (push operation) or gets the rumor from j if j knows the rumor (pull operation). In the synchronous model, the spread time of a rumor is defined as the number of synchronous rounds necessary for all the nodes to know the rumor. In one of the first papers dealing with the push operation only, Frieze and Grimmet [14] proved that if the underlying graph is complete, then asymptotically almost surely the number of rounds is $\log_2(n) + \log(n) + o(\log n)$ where n is the number of nodes of the graph. Further results have been established (see for example [20], [6] and the references herein), the most recent ones resulting from the observation that the rumor spreading time is closely related to the conductance of the graph of the network [15], [16]. Investigations have also been done in different topologies of the network [8], [10], [12], [23], in the presence of link or vertices failures (see [27]), and dynamic graphs [9].

All the above studies assume that all vertices of the network act synchronously. In distributed networks, and in particular in large scale distributed systems, such a strict synchronization is unrealistic. Several authors have recently dropped this assumption by considering an asynchronous model. Boyd et al [26] consider that each node has a clock that goes off at the time of a rate 1 Poisson process. Each time the ring of a node goes off, the push or pull operations are triggered according to the knowledge of the node. Can et al. [18] go a step further by studying rumor spreading time for any graph topology. They show that both the average and guaranteed spreading time are $\Omega(n)$, where n is the number of nodes in the network. Further investigations have been made for different network topologies [24], [13].

This work was partially funded by the French ANR project SocioPlug (ANR-13-INFR-0003), and by the DeSceNt project granted by the Labex CominLabs excellence laboratory (ANR-10-LABX-07-01).

a) Our contributions : In this paper we consider the population protocol model, which turns out to resemble to the discrete-time version of the asynchronous spreading model. This model provides minimalist assumptions on the computational power of the nodes: nodes are finite-state automata, identically programmed, they have no identity, they do not know how numerous they are, and they progress in their computation through random pairwise interactions. Their objective is to ultimately converge to a state from which the sought property can be derived from any node [4]. In this model, the spreading time is defined as the number of interactions needed for all the nodes of the network to learn the rumor. Angluin et al [2] analyze the spreading time of a rumor by only considering the push operation (which they call the one-way epidemic operation), and show that with high probability, a rumor injected at some node requires $O(n \log n)$ interactions to be spread to all the nodes of the network.

In the present paper we go a step further by considering a more general problem namely, that is all the nodes of the network initially receive an input value, and the objective for each node is to learn the maximal value initially received by any node. Note that the rumor spreading problem is a particular instance of this problem when two input values 1 and 0 are considered respectively representing the knowledge and the absence of knowledge of the rumor. We present a thorough analysis of the number of interactions needed for all the nodes to converge to the correct response. Specifically, we study the expectation, variance and an exact formulation of the distribution of the number of interactions needed to propagate a rumor.

This formulation being hardly usable in practice once n becomes too large, a tight bound is derived. This bound is all the more interesting as usual probabilistic inequalities fail to provide relevant results in this case. Finally, we study the asymptotic behavior of the spreading time when the size of the network tends to infinity.

b) Road map: The remainder of this paper is organized as follows. Section II presents the population protocol model. Section III specifies the problem addressed in this work. Analysis of the spreading time is proposed in Section IV, while we study in Section V its asymptotic behavior. We have simulated our protocol to illustrate our theoretical analysis. Finally, Section VI concludes.

II. Population protocols model

In this section, we present the population protocol model, introduced by Angluin et al. [1]. This model describes the behavior of a collection of nodes that interact pairwise. The following definition is from Angluin et al [3]. A population protocol is characterized by a 6-tuple $(Q, \Sigma, Y, \iota, \omega, f)$, over a complete interaction graph linking the set of n nodes, where Q is a finite set of states, Σ is a finite set of input symbols, Y is a finite set of output symbols, $\iota : \Sigma \to Q$ is the input function that determines the initial state of a node, $\omega : Q \to Y$ is the output function that determines the output symbol of a node, and $f : Q \times Q \to Q \times Q$ is the transition function that describes how two nodes interact and update theirs states. Initially all the nodes start with a initial symbol from Σ, and upon interactions with nodes update their state according to the transition function f. Interactions between nodes are orchestrated by a random scheduler: at each discrete time, any two nodes are randomly chosen to interact with a given distribution. Note that its is assumed that the random scheduler is fair, which means that the interactions distribution is such that any possible interaction cannot be avoided forever. The notion of time in population protocols refers to as the successive steps at which interactions occur, while the parallel time refers to as the successive number of steps each node executes [4]. Nodes do not maintain nor use identifiers (nodes are anonymous and cannot determine whether any two interactions have occurred with the same agents or not). However, for ease of presentation the nodes are numbered $1, 2, \ldots, n$. We denote by $C_t^{(i)}$ the state of node i at time t. The stochastic process $C = \{C_t,\ t \geq 0\}$, where $C_t = (C_t^{(1)}, \ldots, C_t^{(n)})$, represents the evolution of the population protocol. The state space of C is thus Q^n and a state of this process is also called a protocol configuration.

III. Spreading the maximum

We consider in this section the following problem. Each site has initially an integer value. At each discrete instant of time, two distinct nodes are successively chosen and they change their value with the maximum value of each node. More precisely, for all nodes a and b, with $a \neq b$, we consider the function f given by

$$f(a, b) = (\max\{a, b\}, \max\{a, b\}) .$$

We want to evaluate the time needed so that all the nodes get the same value.

Let $C = \{C_t,\ t \geq 0\}$ be a discrete-time stochastic process with state space $S = \mathbb{N}^n$. For every $t \geq 0$, the state at time t of the process is denoted by $C_t = (C_t^{(1)}, \ldots, C_t^{(n)})$, where $C_t^{(i)}$ is the integer value of node i at time t. At each instant t, two distinct indexes i and j are successively chosen among the set of nodes $1, \ldots, n$ randomly. We denote by X_t the random variable representing this choice and we suppose that this choice is uniform, i.e we suppose that

$$\mathbb{P}\{X_t = (i, j)\} = \frac{1}{n(n-1)} 1_{\{i \neq j\}}.$$

Once the couple (i, j) is chosen at time t, the process reaches state C_{t+1}, at time $t+1$, given by

$$\begin{aligned} C_{t+1}^{(i)} = C_{t+1}^{(j)} &= \max\{C_t^{(i)}, C_t^{(j)}\} \\ \text{and } C_{t+1}^{(m)} &= C_t^{(m)} \text{ for } i \neq j. \end{aligned}$$

We denote by M the maximum initial value among all the nodes, i.e. $M = \max\{C_0^{(1)}, \ldots, C_0^{(n)}\}$. It is easily checked that for all $t \geq 0$, we have $M = \max\{C_t^{(1)}, \ldots, C_t^{(n)}\}$.

We consider the random variable T_n defined by

$$T_n = \inf\{t \geq 0 \mid C_t^{(i)} = M, \text{ for every } 1, \ldots, n\}.$$

The random variable T_n represents the number of interactions needed for all the nodes in the network to know the maximal value M.

We introduce the discrete-time stochastic process $Y = \{Y_t,\ t \geq 0\}$ with state space $\{1, \ldots, n\}$ defined, for all $t \geq 0$, by

$$Y_t = \left| \left\{ i \mid C_t^{(i)} = M \right\} \right|.$$

The random variable Y_t represents the number of nodes knowing the maximum value M at time t. The stochastic process Y is then a homogeneous Markov chain with transition probability matrix A. The non zero transition probabilities are given, for $i, j = 1, \ldots, n$, by

$$\begin{cases} A_{i,i} &= 1 - \dfrac{2i(n-i)}{n(n-1)}, \\ \\ A_{i,i+1} &= \dfrac{2i(n-i)}{n(n-1)}, \text{ for } i \neq n. \end{cases}$$

Indeed, when $Y_t = i$, in order to get $Y_{t+1} = i + 1$, either the first node must be chosen among the ones with the maximum value (probability i/n) and the second agent must be chosen among the ones with the non maximum value (probability $(n-i)/(n-1)$) or the first node must be chosen among the ones with the non maximum value (probability $(n-i)/n$) and the second node must be chosen among the ones with the non maximum value (probability $i/(n-1)$).

The states $1, \ldots, n-1$ of Y are transient and state n is absorbing. The random variable T_n can then be written as

$$T_n = \inf\{t \geq 0 \mid Y_t = n\}.$$

It is well-known, see for instance [25], that the distribution of T_n is given, for every $k \geq 0$, by

$$\mathbb{P}\{T_n > k\} = \alpha Q^k \mathbb{1}, \tag{1}$$

where α is the row vector containing the initial probabilities of states $1, \ldots, n-1$, that is $\alpha_i = \mathbb{P}\{Y_0 = i\}$, Q is the sub-matrix obtained from A by deleting the row and the column corresponding to absorbing state n and $\mathbb{1}$ is the column vector of dimension $n-1$ with all its entries equal to 1.

For $i = 0, \ldots, n$, we introduce the notation

$$p_i = \frac{2i(n-i)}{n(n-1)}$$

and we denote by H_k the harmonic series defined by $H_0 = 0$ and $H_k = \sum_{\ell=1}^{k} 1/\ell$, for $k \geq 1$. Note that, for every $i = 0, \ldots, n$, we have $p_i = p_{n-i}$.

If we denote by S_i, for $i = 1, \ldots, n-1$, the total time spent by the Markov chain Y in state i, then conditionally on the event $Y_0 = i$, S_ℓ has a geometric distribution with parameter p_ℓ, for $\ell = i, \ldots, n-1$ and in this case, we have $T_n = S_i + \cdots + S_{n-1}$. It follows that

$$\mathbb{P}\{T_n > k \mid Y_0 = i\} = \mathbb{P}\{S_i + \cdots + S_{n-1} > k\},$$

which means that $\mathbb{P}\{T_n > k \mid Y_0 = i\}$ is decreasing with i and in particular that

$$\mathbb{P}\{T_n > k \mid Y_0 = i\} \leq \mathbb{P}\{T_n > k \mid Y_0 = 1\}. \tag{2}$$

IV. Analysis of the spreading time

In the following we study the expectation and the variance of T_n, the number of interactions needed for all the nodes in the network to know the maximal value M. We then provide an explicit expression of the distribution of T_n, and then a bound and an equivalent for the explicit distribution of T_n.

A. Expectation and variance of T_n

The mean time $\mathbb{E}(T_n)$ needed so that all the nodes get the same value is then given by

$$\mathbb{E}(T_n) = \alpha(I - Q)^{-1}\mathbb{1}, \tag{3}$$

where I is the identity matrix. This expectation can also be written as

$$\mathbb{E}(T_n) = \sum_{i=1}^{n-1} \alpha_i \mathbb{E}(T_n \mid Y_0 = i).$$

This conditional expectations are given by the following theorem.

Theorem *1:* For every $n \geq 1$ and $i = 1, \ldots, n$, we have

$$\mathbb{E}(T_n \mid Y_0 = i) = \frac{(n-1)(H_{n-1} + H_{n-i} - H_{i-1})}{2}.$$

Proof: If $Y_0 = n$, which means that all the nodes start with same values, then we have $T_n = 0$ and so $\mathbb{E}(T_n \mid Y_0 = n) = 0$. For $i = 1, \ldots, n-1$ we have

$$\begin{aligned} \mathbb{E}(T_n \mid Y_0 = i) &= \sum_{\ell=i}^{n-1} \mathbb{E}(S_\ell) \\ &= \sum_{\ell=i}^{n-1} \frac{1}{p_\ell} \\ &= \frac{n(n-1)}{2} \sum_{\ell=i}^{n-1} \frac{1}{\ell(n-\ell)} \\ &= \frac{n-1}{2} \sum_{\ell=i}^{n-1} \left(\frac{1}{\ell} + \frac{1}{n-\ell} \right) \\ &= \frac{(n-1)(H_{n-1} + H_{n-i} - H_{i-1})}{2}, \end{aligned}$$

which completes the proof. ∎

In particular, when the maximum value is initially unique, i.e. when $Y_0 = 1$ with probability 1, we have $\alpha_1 = 1$ and thus

$$\mathbb{E}(T_n) = \mathbb{E}(T_n \mid Y_0 = 1) = (n-1)H_{n-1} \underset{n \to \infty}{\sim} n \ln(n).$$

More generally, from Relation (2), we have

$$\mathbb{E}(T_n) \leq \mathbb{E}(T_n \mid Y_0 = 1) = (n-1)H_{n-1} \underset{n \to \infty}{\sim} n \ln(n).$$

The variance of T_n is obtained similarly.

Theorem *2:* For every $n \geq 1$ and $i = 1, \ldots, n$, we have

$$\begin{aligned} \operatorname{Var}(T_n \mid Y_0 = i) = &\frac{(n-1)^2}{4} \left(\sum_{\ell=i}^{n-1} \frac{1}{\ell^2} + \sum_{\ell=1}^{n-i} \frac{1}{\ell^2} \right) \\ &- \frac{\mathbb{E}(T_n \mid Y_0 = i)}{n}. \end{aligned}$$

Proof: If $Y_0 = n$, which means that all the nodes start with the same values, then we have $T_n = 0$ and thus $\text{Var}(T_n \mid Y_0 = n) = 0$. For $i = 1, \dots, n-1$ we have, using the independence of the S_ℓ,

$$\begin{aligned}
\text{Var}(T_n \mid Y_0 = i) &= \sum_{\ell=i}^{n-1} \text{Var}(S_\ell) = \sum_{\ell=i}^{n-1} \frac{1-p_\ell}{p_\ell^2} \\
&= \sum_{\ell=i}^{n-1} \frac{1}{p_\ell^2} - \sum_{\ell=i}^{n-1} \frac{1}{p_\ell} \\
&= \frac{n^2(n-1)^2}{4} \sum_{\ell=i}^{n-1} \frac{1}{\ell^2(n-\ell)^2} - \frac{n(n-1)}{2} \sum_{\ell=i}^{n-1} \frac{1}{\ell(n-\ell)} \\
&= \frac{(n-1)^2}{4} \sum_{\ell=i}^{n-1} \left(\frac{1}{\ell} + \frac{1}{n-\ell}\right)^2 - \frac{n(n-1)}{2} \sum_{\ell=i}^{n-1} \frac{1}{\ell(n-\ell)} \\
&= \frac{(n-1)^2}{4} \sum_{\ell=i}^{n-1} \left(\frac{1}{\ell^2} + \frac{1}{(n-\ell)^2}\right) - \frac{n-1}{2} \sum_{\ell=i}^{n-1} \frac{1}{\ell(n-\ell)} \\
&= \frac{(n-1)^2}{4} \sum_{\ell=i}^{n-1} \left(\frac{1}{\ell^2} + \frac{1}{(n-\ell)^2}\right) - \frac{\mathbb{E}(T_n \mid Y_0 = i)}{n} \\
&= \frac{(n-1)^2}{4} \left(\sum_{\ell=i}^{n-1} \frac{1}{\ell^2} + \sum_{\ell=1}^{n-i} \frac{1}{\ell^2}\right) - \frac{\mathbb{E}(T_n \mid Y_0 = i)}{n},
\end{aligned}$$

which completes the proof. ■

In particular, when the maximum value is initially unique, i.e. when $Y_0 = 1$ with probability 1, we have $\alpha_1 = 1$ and thus

$$\begin{aligned}
\text{Var}(T_n) &= \text{Var}(T_n \mid Y_0 = 1) \\
&= \frac{(n-1)^2}{2} \sum_{\ell=1}^{n-1} \frac{1}{\ell^2} - \frac{n-1}{n} H_{n-1} \underset{n \to \infty}{\sim} \frac{\pi^2 n^2}{12}.
\end{aligned}$$

More generally, from Theorem 2, we have

$$\begin{aligned}
\text{Var}(T_n \mid Y_0 = i) &\le \frac{(n-1)^2}{4} \left(\sum_{\ell=i}^{n-1} \frac{1}{\ell^2} + \sum_{\ell=1}^{n-i} \frac{1}{\ell^2}\right) \\
&\le \frac{(n-1)^2}{2} \sum_{\ell=1}^{n-1} \frac{1}{\ell^2} \le \frac{\pi^2 n^2}{12}.
\end{aligned}$$

It follows that

$$\text{Var}(T_n) = \sum_{i=1}^{n-1} \alpha_i \text{Var}(T_n \mid Y_0 = i) \le \frac{\pi^2 n^2}{12}.$$

B. Explicit expression of the distribution of T_n

The distribution of T_n, for $n \ge 2$, which is given by Relation (1) can be easily computed as follows. Let $V(k) = (V_1(k), \dots, V_{n-1}(k))$ be the column vector defined by $V_i(k) = \mathbb{P}\{T_n > k \mid Y_0 = i\}$. According to Relation (1), we have $V(k) = Q^k \mathbb{1}$. Since $V(0) = \mathbb{1}$, writing $V(k) = QV(k-1)$ for $k \ge 1$, we get for $i = 1, \dots, n-2$,

$$\begin{cases} V_i(k) = (1-p_i)\, V_i(k-1) + p_i V_{i+1}(k-1), \\ V_{n-1}(k) = (1-p_{n-1})\, V_{n-1}(k-1). \end{cases} \tag{4}$$

Recall that we have $p_i = 2i(n-i)/(n(n-1))$. This recursion can be easily computed since we have, for $k \ge 0$,

$$V_{n-1}(k) = (1-p_{n-1})^k = \left(1 - \frac{2}{n}\right)^k. \tag{5}$$

In the next theorem, we derive from recursion (4) an explicit expression of the distribution of T_n.

Theorem *3:* For every $n \ge 1$, $k \ge 0$ and $i = 1, \dots, n-1$, we have

$$\begin{aligned}
&\mathbb{P}\{T_n > k \mid Y_0 = n-i\} \\
&\qquad = \sum_{j=1}^{\lfloor n/2 \rfloor} \left(c_{i,j}(1-p_j) + k d_{i,j}\right)(1-p_j)^{k-1},
\end{aligned}$$

where the coefficients $c_{i,j}$ and $d_{i,j}$, which do not depend on k, are given, for $j = 1, \dots, n-1$, by

$$c_{1,j} = 1_{\{j=1\}} \text{ and } d_{1,j} = 0$$

and for $i \in \{2, \dots, n-1\}$ by

$$\begin{aligned}
c_{i,j} &= \frac{p_i c_{i-1,j}}{p_i - p_j} - \frac{p_i d_{i-1,j}}{(p_i - p_j)^2} && \text{for } i \ne j, n-j, \\
d_{i,j} &= \frac{p_i d_{i-1,j}}{p_i - p_j} && \text{for } i \ne j, n-j, \\
c_{i,i} &= 1 - \sum_{j=1, j \ne i}^{n/2} c_{i,j} && \text{for } i \le n/2, \\
c_{i,n-i} &= 1 - \sum_{j=1, j \ne n-i}^{n/2} c_{i,j} && \text{for } i > n/2, \\
d_{i,i} &= p_i c_{i-1,i} && \text{for } i \le n/2, \\
d_{i,n-i} &= p_i c_{i-1,n-i} && \text{for } i > n/2.
\end{aligned}$$

Proof: See appendix ■

C. Bounds of the distribution of T_n

The exact expression of the distribution of T_n presented earlier is hardly usable in practice, and computation using formula (4) may take a long time for large values of n. To overcome this problem, we propose in this section a bound and an equivalent for the quantity $\mathbb{P}\{T_n > k \mid Y_0 = i\}$ derived from the recursive formula (4).

Theorem *4:* For all $n \ge 2$ and $k \ge 1$ we have

$$\mathbb{P}\{T_n > k \mid Y_0 = 1\} \le \left(1 + \frac{2k(n-2)^2}{n}\right)\left(1 - \frac{2}{n}\right)^{k-1},$$

$$\mathbb{P}\{T_n > k \mid Y_0 = 1\} \underset{k \to \infty}{\sim} \left(1 + \frac{2k(n-2)^2}{n}\right)\left(1 - \frac{2}{n}\right)^{k-1}$$

and for $i = 2, \dots, n-1$ and $k \ge 0$,

$$\mathbb{P}\{T_n > k \mid Y_0 = i\} \le \frac{(n-i)(n-2)}{i-1}\left(1 - \frac{2}{n}\right)^k,$$

$$\mathbb{P}\{T_n > k \mid Y_0 = i\} \underset{k \to \infty}{\sim} \frac{(n-i)(n-2)}{i-1}\left(1 - \frac{2}{n}\right)^k.$$

Moreover, we have

$$\mathbb{P}\{T_n > k\} \leq \mathbb{P}\{T_n > k \mid Y_0 = 1\}.$$

Proof: The result is trivial for $n = 2$ since in this case we have $T_2 = 1$. We thus suppose that $n \geq 3$. Note that by definition of p_i we have $p_i = p_{n-i}$. Consider the sequence b_i defined for $i = 1, \ldots, n-2$, by

$$b_1 = 1 \text{ and } b_i = \frac{p_i b_{i-1}}{p_i - p_1}, \text{ for } i = 2, \ldots, n-2.$$

Observing that

$$b_i = \frac{i(n-i)b_{i-1}}{(i-1)(n-i-1)},$$

it is easily checked by recurrence that for $i = 1, \ldots, n-2$, we have

$$b_i = \frac{i(n-2)}{n-i-1}.$$

We show now by recurrence that for all $i = 1, \ldots, n-2$, we have

$$V_{n-i}(k) \leq b_i \left(1 - p_1\right)^k, \text{ for all } k \geq 0$$
$$\text{and } V_{n-i}(k) \underset{k \longrightarrow \infty}{\sim} b_i \left(1 - p_1\right)^k.$$

Both results are true for $i = 1$ since $V_{n-1}(k) = (1-p_{n-1})^k = (1 - p_1)^k$. Suppose now that these results are true for a fixed integer i with $1 \leq i \leq n-3$. From Relations (4), we have

$$\begin{aligned} &V_{n-i-1}(k) \\ &\quad= (1 - p_{n-i-1})V_{n-i-1}(k-1) + p_{n-i-1}V_{n-i}(k-1) \\ &\quad= (1 - p_{i+1})V_{n-i-1}(k-1) + p_{i+1}V_{n-i}(k-1). \end{aligned}$$

Using the recurrence hypothesis, we obtain, for what concerns the inequality,

$$V_{n-i-1}(k) \leq (1 - p_{i+1})V_{n-i-1}(k-1) + p_{i+1}b_i \left(1 - p_1\right)^{k-1}.$$

Expanding this inequality and using the fact that $V_{n-i-1}(0) = 1$, this leads to

$$\begin{aligned} &V_{n-i-1}(k) \\ &\quad\leq (1 - p_{i+1})^k + p_{i+1}b_i \sum_{j=0}^{k-1} (1 - p_{i+1})^j (1 - p_1)^{k-1-j} \\ &\quad= (1 - p_{i+1})^k + p_{i+1}b_i \frac{(1 - p_1)^k - (1 - p_{i+1})^k}{p_{i+1} - p_1} \\ &\quad= (1 - p_{i+1})^k + b_{i+1}\left((1 - p_1)^k - (1 - p_{i+1})^k\right) \\ &\quad= (1 - b_{i+1})\left(1 - p_{i+1}\right)^k + b_{i+1}(1 - p_1)^k. \end{aligned}$$

Since $b_{i+1} \geq 1$, we get

$$V_{n-i-1}(k) \leq b_{i+1}(1 - p_1)^k$$

In the same way, using a similar calculus, we obtain

$$V_{n-i-1}(k) \underset{k \longrightarrow \infty}{\sim} (1 - b_{i+1})\left(1 - p_{i+1}\right)^k + b_{i+1}(1 - p_1)^k.$$

Since $p_{i+1} > p_1$, we also get

$$V_{n-i-1}(k) \underset{k \longrightarrow \infty}{\sim} b_{i+1}(1 - p_1)^k.$$

We thus have shown that for all $i = 1, \ldots, n-2$, we have

$$V_{n-i}(k) \leq b_i \left(1 - p_1\right)^k, \text{ for all } k \geq 0$$
$$\text{and } V_{n-i}(k) \underset{k \longrightarrow \infty}{\sim} b_i \left(1 - p_1\right)^k.$$

In particular, for $i = n-2$ we obtain

$$V_2(k) \leq b_{n-2} \left(1 - p_1\right)^k, \text{ for all } k \geq 0$$
$$\text{and } V_2(k) \underset{k \longrightarrow \infty}{\sim} b_{n-2} \left(1 - p_1\right)^k.$$

Consider now the term $V_1(k)$. From Relations (4) and using the previous inequality, we have

$$\begin{aligned} V_1(k) &= (1 - p_1)V_1(k-1) + p_1V_2(k-1) \\ &\leq (1 - p_1)V_1(k-1) + p_1b_{n-2} \left(1 - p_1\right)^{k-1}. \end{aligned}$$

Expanding this inequality and using the fact that $V_1(0) = 1$, this leads to

$$\begin{aligned} V_1(k) &\leq (1 - p_1)^k + p_1b_{n-2} \sum_{j=0}^{k-1} (1 - p_1)^j (1 - p_1)^{k-1-j} \\ &= (1 - p_1)^k + p_1b_{n-2}k(1 - p_1)^{k-1} \\ &= (1 - p_1 + kp_1b_{n-2}) \left(1 - p_1\right)^{k-1} \\ &\leq (1 + kp_1b_{n-2}) \left(1 - p_1\right)^{k-1}, \end{aligned}$$

which gives

$$V_1(k) \leq \left(1 + \frac{2k(n-2)^2}{n}\right) \left(1 - \frac{2}{n}\right)^{k-1}.$$

In the same way, using a similar calculus, we obtain

$$V_1(k) \underset{k \longrightarrow \infty}{\sim} \left(1 + \frac{2k(n-2)^2}{n}\right) \left(1 - \frac{2}{n}\right)^{k-1}.$$

Finally, since $\mathbb{P}\{T_n > k \mid Y_0 = i\}$ is decreasing with i, we have

$$\begin{aligned} \mathbb{P}\{T_n > k\} &= \sum_{i=1}^{n-1} \mathbb{P}\{T_n > k \mid Y_0 = i\}\mathbb{P}\{Y_0 = i\} \\ &\leq \mathbb{P}\{T_n > k \mid Y_0 = 1\}, \end{aligned}$$

which completes the proof. ∎

The bound established in Theorem 4 is all the more interesting as usual probabilistic inequalities fail to provide relevant results in this particular case. For example, Markov inequality leads for all real number $c \geq 1$ to

$$\mathbb{P}\{T_n \geq c\mathbb{E}(T_n)\} \leq \frac{1}{c},$$

and Bienaym-Tchebychev inequality leads for all real number $x > 0$ to

$$\mathbb{P}\{|T_n - \mathbb{E}(T_n)| \geq x\} \leq \frac{\pi^2 n^2}{12x^2}.$$

The author of [19] provides a bound, based on Chernoff inequality, for the tail probabilities of the sum of independent, but not necessarily identically distributed, geometric random variables. In the particular case of our protocol computing the maximum, this leads to the following result.

Theorem 5: For all $n \geq 3$ and for all real number $c \geq 1$, we have

$$\mathbb{P}(T_n > c\mathbb{E}(T_n)) \leq \frac{1}{c}\left(1 - \frac{2}{n}\right)^{(c-1-\ln c)(n-1)H_{n-1}}.$$

The right-hand side term is equal to 1 when $c = 1$.

Proof: We have already shown that

$$\mathbb{P}(T_n > c\mathbb{E}(T_n)) \leq \mathbb{P}(T_n > c\mathbb{E}(T_n) \mid Y_0 = 1\}.$$

The upper bound is then an application of Theorem 2.3 of [19], and it is clearly equal to 1 when $c = 1$. ■

Applying Theorem 4 at point $k = \lfloor c\mathbb{E}(T_n)\rfloor$, we obtain

$$\begin{aligned} \mathbb{P}(T_n > c\mathbb{E}(T_n)) &\leq \left(1 + \frac{2\lfloor c\mathbb{E}(T_n)\rfloor(n-2)^2}{n}\right) \\ &\quad \times \left(1 - \frac{2}{n}\right)^{\lfloor c\mathbb{E}(T_n)\rfloor - 1} \\ &\leq \left(1 + \frac{2c\mathbb{E}(T_n)(n-2)^2}{n}\right) \\ &\quad \times \left(1 - \frac{2}{n}\right)^{c\mathbb{E}(T_n) - 2}. \end{aligned}$$

From now on we denote this bound by $f(c, n)$ and in the same way, we denote by $g(c, n)$ the bound of $\mathbb{P}(T_n > c\mathbb{E}(T_n))$ derived from Theorem 5. We then have, for $n \geq 3$ and $c \geq 1$,

$$\begin{aligned} f(c,n) &= \left(1 + \frac{2c(n-1)H_{n-1}(n-2)^2}{n}\right) \\ &\quad \times \left(1 - \frac{2}{n}\right)^{c(n-1)H_{n-1} - 2} \\ g(c,n) &= \frac{1}{c}\left(1 - \frac{2}{n}\right)^{(c-1-\ln(c))(n-1)H_{n-1}}. \end{aligned}$$

We also introduce the notation

$$e(c, n) = \mathbb{P}(T_n > c\mathbb{E}(T_n)).$$

Theorem 6: For every $n \geq 3$, there exists a unique $c^* \geq 1$ such that $f(c^*, n) = g(c^*, n)$ and we have

$$\begin{cases} f(c,n) > g(c,n) & \text{for all } 1 \leq c < c^* \\ f(c,n) < g(c,n) & \text{for all } c > c^*. \end{cases} \tag{6}$$

Furthermore,

$$\lim_{c \longrightarrow \infty} \frac{f(c,n)}{g(c,n)} = 0.$$

Proof: See appendix. ■

The graphs on Figures 1, 2 and 3 illustrate the behavior of the bounds $f(c, n)$ and $g(c, n)$, depending on c and for different values of n, compared to the real distribution of T_n at point $c\mathbb{E}(T_n)$, i.e. to $e(c, n) = \mathbb{P}\{T_n > \mathbb{E}(T_n)\}$. The bound $f(c, n)$ that we provided in Theorem 4 clearly shows better accuracy than the Chernoff bound $g(c, n)$ provided in [19] above the threshold c^* introduced in Theorem 6. Furthermore, this threshold seems to decrease to 1 as n tends to infinity, as can be seen on Figure 4.

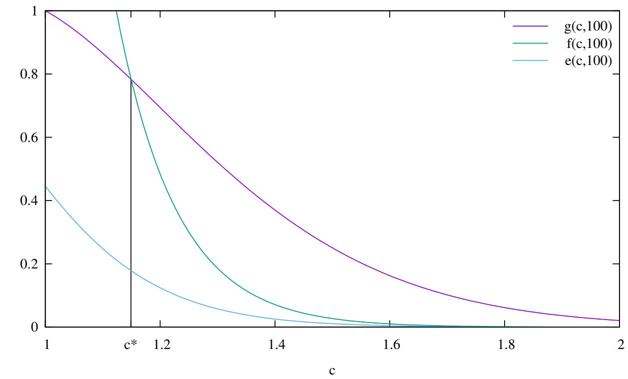

Fig. 1. Bounds $f(c, n)$ and $g(c, n)$ beside the real value of $\mathbb{P}(T_n > c\mathbb{E}(T_n)) = e(c, n)$ for $n = 100$, as functions of c. In this case, we have $c^* = 1.14641$.

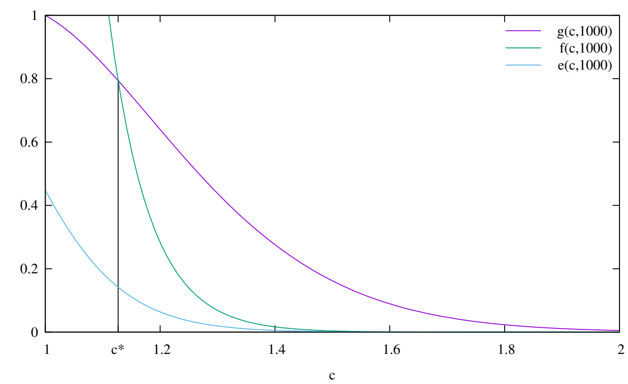

Fig. 2. Bounds $f(c, n)$ and $g(c, n)$ beside the real value of $\mathbb{P}(T_n > c\mathbb{E}(T_n)) = e(c, n)$ for $n = 1000$, as functions of c. In this case, we have $c^* = 1.12673$.

V. Asymptotic analysis of the distribution of T_n

We analyze in this section the behavior of the complementary distribution of T_n at point $c\mathbb{E}(T_n)$ when n tends to

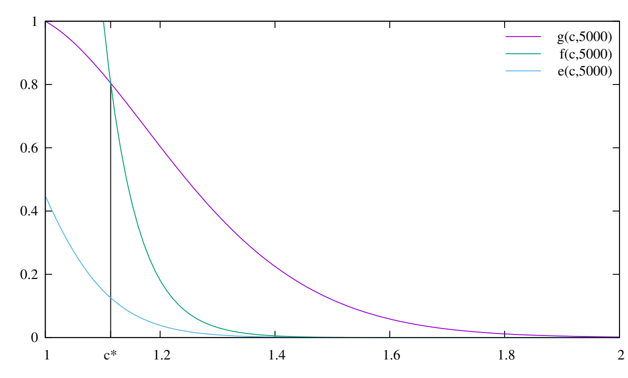

Fig. 3. Bounds $f(c, n)$ and $g(c, n)$ beside the real value of $\mathbb{P}(T_n > c\mathbb{E}(T_n)) = e(c, n)$ for $n = 5000$, as functions of c. In this case, we have $c^* = 1.11385$.

n	10	10^2	10^3	10^4	10^5	10^6	10^7
c^*	1.09	1.15	1.13	1.11	1.10	1.09	1.08

Fig. 4. Approximate values of c^* for different network sizes n.

infinity, depending on the value of c.

We prove in the following corollary that the bounds $f(c,n)$ and $g(c,n)$, obtained from Theorem 4 and Theorem 5 respectively with $k = c\mathbb{E}(T_n)$, both tend to 0 when n goes to infinity.

Corollary *7:* For all real number $c > 1$, we have

$$\lim_{n\longrightarrow\infty} f(c,n) = 0 \text{ and } \lim_{n\longrightarrow\infty} g(c,n) = 0.$$

Proof: For all $x \in [0,1)$, we have $\ln(1-x) \leq -x$. Applying this property to the bound $f(c,n)$ leads to

$$\begin{aligned} f(c,n) &\leq \left(1 + \frac{2c(n-1)H_{n-1}(n-2)^2}{n}\right) \\ &\quad \times e^{-2(c(n-1)H_{n-1}-2)/n} \\ &\leq \left(1 + 2c(n-2)^2 H_{n-1}\right) e^{-2(c(n-1)H_{n-1}-2)/n}. \end{aligned}$$

Since $\ln(n) \leq H_{n-1} \leq 1 + \ln(n-1)$, we get

$$\begin{aligned} f(c,n) &\leq \left(1 + 2c(n-2)^2(1+\ln(n-1)\right) \\ &\quad \times e^{-2(c(n-1)\ln(n)-2)/n} \\ &= \left(1 + 2c(n-2)^2(1+\ln(n-1))\right) e^{-2c\ln(n)} \\ &\quad \times e^{2(c\ln(n)+2)/n}. \end{aligned}$$

For $x \geq 0$, the function $u(x) = e^{2(c\ln(x)+2)/x}$ satisfies $u(x) \leq \exp\left(2c/e^{(c-2)/c}\right)$, so we obtain

$$f(c,n) \leq \frac{1 + 2c(n-2)^2(1+\ln(n-1))}{n^{2c}} \exp\left(2c/e^{(c-2)/c}\right).$$

The fact that $c > 1$ implies that this last term tends to 0 when $n \longrightarrow \infty$. Concerning the bound $g(c,n)$, we have

$$\begin{aligned} g(c,n) &= \frac{1}{c}\left(1 - \frac{2}{n}\right)^{(c-1-\ln(c))(n-1)H_{n-1}} \\ &= \frac{1}{c} e^{(c-1-\ln(c))(n-1)H_{n-1}\ln(1-2/n)} \\ &\leq \frac{1}{c} e^{-2(c-1-\ln(c))(n-1)H_{n-1}/n}, \end{aligned}$$

which tends to 0 when n tends to infinity, since $c - 1 - \ln(c)$ is positive for $c > 1$. ∎

Theorem *8:* For all real $c \geq 0$, we have

$$\lim_{n\to+\infty} \mathbb{P}\{T_n > c\mathbb{E}(T_n)\} = \begin{cases} 0 & \text{if } c > 1 \\ 1 & \text{if } c < 1. \end{cases}$$

Proof: From Corollary 7, both bounds $f(c,n)$ and $g(c,n)$ of $\mathbb{P}\{T_n > c\mathbb{E}(T_n)\}$ tend to 0 when n tends to infinity, so using either $f(c,n)$ or $g(c,n)$ we deduce that

$$\lim_{n\longrightarrow\infty} \mathbb{P}\{T_n > c\mathbb{E}(T_n)\} = 0 \text{ for all } c > 1.$$

In the case where $c < 1$, Theorem 3.1 of [19] leads to

$$\begin{aligned} \mathbb{P}\{T_n > c\mathbb{E}(T_n)\} &\geq 1 - e^{-2(n-1)H_{n-1}(c-1-\ln(c))/n} \\ &\geq 1 - e^{-2(n-1)\ln(n)(c-1-\ln(c))/n}. \end{aligned}$$

Since $c - 1 - \ln(c) > 0$ for all $c \in [0,1)$, the right-hand side term of this inequality tends to 1 when $n \longrightarrow \infty$. Thus, $\lim_{n\longrightarrow\infty} \mathbb{P}\{T_n > c\mathbb{E}(T_n)\} = 1$ when $c < 1$. ∎

The results established previously don't allow us to figure out neither the existence of $\lim_{n\longrightarrow\infty} \mathbb{P}\{T_n > c\mathbb{E}(T_n)\}$ when $c = 1$, nor its value. However, numerical results give us a glimpse of its limiting behavior.

In Figure 5, we show the probability $\mathbb{P}\{T_n > \mathbb{E}(T_n)\}$ for different values of n. The oscillations of this probability with n are due to the fact T_n is a discrete random variable and $\mathbb{E}(T_n)$ is not an integer. That is why we propose in this figure a smoothing of this probability using the sequence

$$s_n = (1-a_n)\mathbb{P}\{T_n > \mathbb{E}(T_n)\} + a_n\mathbb{P}\{T_n > \mathbb{E}(T_n)+1\},$$

where a_n is the fractional part of $\mathbb{E}(T_n)$, that is $a_n = \mathbb{E}(T_n) - \lfloor\mathbb{E}(T_n)\rfloor$. Since $a_n \in [0,1]$, we have

$$\mathbb{P}\{T_n > \mathbb{E}(T_n)+1\} \leq s_n \leq \mathbb{P}\{T_n > \mathbb{E}(T_n)\},$$

that is why we also show in this figure the probability $\mathbb{P}\{T_n > \mathbb{E}(T_n)+1\}$. We also had verified that the sequence (s_n) is increasing until $n = 20000$. This fact and this figure suggests the following result proposed as a conjecture.

Conjecture : $\lim_{n\longrightarrow\infty} \mathbb{P}\{T_n > \mathbb{E}(T_n)\}$ exists and ≈ 0.4484.

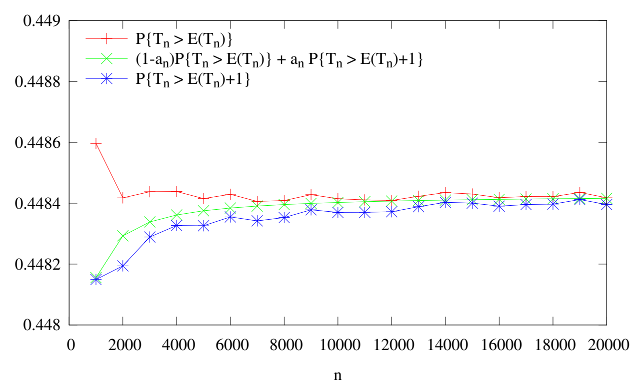

Fig. 5. $\mathbb{P}\{T_n > \mathbb{E}(T_n)\}$ as a function of n and its smoothing obtained with $a_n = \mathbb{E}(T_n) - \lfloor\mathbb{E}(T_n)\rfloor$.

Figure 6 shows the proportion of nodes informed by rumor as a function of the parallel time. Recall that the parallel time refers to as the successive number of steps each node executes [4]. Initially, a single node is informed of the rumor. This figure illustrates our analysis. For instance, with probability 1

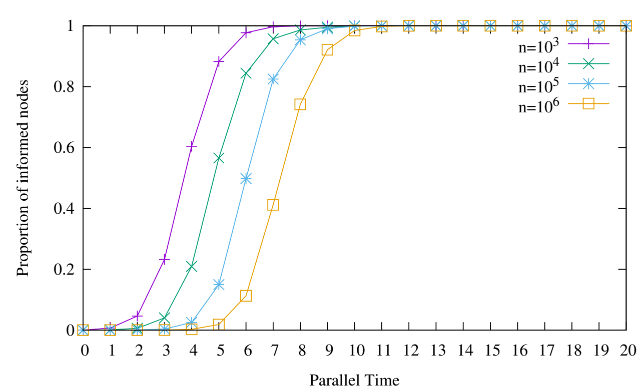

Fig. 6. Proportion of informed nodes as a function of parallel time

one thousand nodes (resp. one million nodes) learn the rumor after no more 7 (resp 11) interactions for each of them. The complexity in space (number of memory bits) is in $O(1)$.

VI. Conclusion

In this paper we have provided a thorough analysis of the rumor spreading time in the population protocol model. Providing such a precise analysis is a step towards the design of more complex functionality achieved by combining simple population protocols [22], [2]. Indeed, an important feature of population protocols is that they do not halt. Nodes can never know whether their computation is completed and thus nodes forever interact with their neighbors while their outputs stabilize to the desired value (e.g. the maximal value of any node of the network). By precisely characterizing, for each protocol of interest, with any high probability, the number of interactions each node must execute to converge to the desired value, each node can on its own, decide the time from which the current protocol has stabilized and start the parallel of sequential executions of the next ones.

References

[1] Dana Angluin, James Aspnes, Zoë Diamadi, Michael J. Fischer, and René Peralta. Computation in networks of passively mobile finite-state sensors. *Distributed Computing*, 18(4):235–253, 2006.

[2] Dana Angluin, James Aspnes, and David Eisenstat. Fast computation by population protocols with a leader. *Distributed Computing*, 21(2):183–199, 2008.

[3] Dana Angluin, James Aspnes, David Eisenstat, and Eric Ruppert. The computational power of population protocols. *Distributed Computing*, 20(4):279–304, 2007.

[4] James Aspnes and Eric Ruppert. An introduction to population protocols. *Bulletin of the European Association for Theoretical Computer Science, Distributed Computing Column*, 93:98–117, 2007.

[5] Noam Berger, Christian Borgs, Jennifer T. Chayes, and Amin Saberi. On the spread of viruses on the internet. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

[6] Marin Bertier, Yann Busnel, and Anne-Marie Kermarrec. On gossip and populations. In *Proceedings of the International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, 2009.

[7] Keren Censor-Hillel, Bernhard Haeupler, Jonathan Kelner, and Petar Maymounkov. Global computation in a poorly connected world: Fast rumor spreading with no dependence on conductance. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, 2012.

[8] Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. Rumor spreading in social networks. *Theoretical Computer Science*, 412(24):2602–2610, 2011.

[9] Andrea Clementi, Pierluigi Crescenzi, Carola Doerr, Pierre Fraigniaud, Francesco Pasquale, and Riccardo Silvestri. Rumor spreading in random evolving graphs. *Random structures and Algorithms*, 48(2):290–312, 2015.

[10] Sebastian Daum, Fabian Kuhn, and Yannic Maus. Rumor spreading with bounded indegree. In *Proceedings of the International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, 2016.

[11] Alan Demers, Mark Gealy, Dan Greene, Carl Hauser, Wes Irish, John Larson, Scott Shenker, Howard Sturgis, Dand Swinehart, and Doug Terry. Epidemic algorithms for replicated datbase maintenance. In *Proceedings of the ACM Syposium on Principles of Distributed Systems (PODC)*, 1987.

[12] Nicolaos Fountoulakis and Konstantinos Panagiotou. Rumor spreading on random regular graphs and expanders. *Random Structures and Algorithms*.

[13] Nicolaos Fountoulakis, Konstantinos Panagiotou, and Thomas Sauerwald. Ultra-fast rumor spreading in social networks.

[14] Alan Frieze and Geoffrey Grimmet. The shortest-path problem for graphs with random arc-lengths. *Discrete Applied Mathematics*, 10(1):57–77, 85.

[15] George Giakkoupis. Tight bounds for rumor spreading in graphs of a given conductance. In *Proceedings of the International Symposium on Theoretical Aspects of Computer Science (STACS)*, 2011.

[16] George Giakkoupis. Tight bounds for rumor spreading with vertex expansion. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2014.

[17] Mor Harchol-Balter, Tom Leighton, and Daniel Lewin. Resource discovery in distributed networks. In *Proceedings of the ACM Syposium on Principles of Distributed Systems (PODC)*, 1999.

[18] Acan Huseyin, Collevecchio Andrea, Mehrabian Abbas, and Wormald Nick. On the push&pull protocol for rumour spreading. 2015.

[19] Svante Janson. Tail bounds for sums of geometric and exponential variables. Technical report. `http://www2.math.uu.se/~svante/papers/sjN14.pdf`

[20] R. Karp, C. Schindelhauer, S. Shenker, and B. Vocking. Randomized rumor spreading. In *Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS)*, 2000.

[21] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *Proceedings of the Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2003.

[22] Othon Michail and Paul Spirakis. Terminating population protocols via some minimal global knowledge assumptions. *Journal of Parallel and Distributed Computing*, 81:1–10, 2015.

[23] Konstantinos Panagiotou, Xavier Perez-Gimenez, Thomas Sauerwald, and Hé Sun. Randomized rumor spreading: the effect of the network topology. *Combinatorics, Probability and Computing*, 24(2):457–479, 2015.

[24] Konstantinos Panagiotou and Leo Speidel. Asynchronous rumor spreading on random graphs. *Algorithmica*, 2016.

[25] Bruno Sericola. *Markov Chains. Theory, Algorithms and Applications*. Applied stochastic methods series. WILEY, 2013.

[26] Boyd Stephen, Ghosh Arpita, Prabhakar Balaji, and Shah Devavrat. Randomized gossip algorithms. *IEEE/ACM Transactions on Networking*, 14:2508–2530, 2006.

[27] P. Raghavan E. Upfal U. Feige, D. Peleg. Randomized broadcast in networks. *Random Structures and Algorithms*, 1(4):447–460, 1990.

Appendix

We give in this appendix the proofs of Theorem 3 and Theorem 6. In order to prove Theorem 3, we first need the following Lemma.

Lemma *9:* Let $N \geq 1$, $a \in (0,1)$, $b_1, \ldots, b_N \in (0,1)$, $c_1, \ldots, c_N \in \mathbb{R}$ and $d_1, \ldots, d_N \in \mathbb{R}$, with the condition,

$$\text{for every } j = 1, \ldots, N, \;\; d_j = 0 \text{ if } b_j = a.$$

Then the sequence $(u_k)_{k \geq 0}$ defined by

$$u_0 = 1 \text{ and } u_{k+1} = a u_k + \sum_{j=1}^{N} (c_j b_j + k d_j) b_j^{k-1}, \; k \geq 0 \quad (7)$$

satisfies

$$u_k = \left(1 - \sum_{j=1}^{N} \theta_j 1_{\{b_j \neq a\}}\right) a^k + \sum_{j=1}^{N} \left((\theta_j b_j + k\gamma_j) 1_{\{b_j \neq a\}} + k c_j 1_{\{b_j = a\}} \right) b_j^{k-1}, \quad (8)$$

where

$$\theta_j = \frac{c_j}{b_j - a} - \frac{d_j}{(b_j - a)^2} \text{ and } \gamma_j = \frac{d_j}{b_j - a}.$$

Proof: We prove this lemma by recurrence. For $k = 0$, Relation (8) gives $u_0 = 1$. We introduce the notation $\alpha = \left(1 - \sum_{j=1}^{N} \theta_j 1_{\{b_j \neq a\}}\right)$ and $f_j(k) = \theta_j b_j + k\gamma_j$. Relation (8) can then be rewritten as

$$u_k = \alpha a^k + \sum_{j=1}^{N} \left[f_j(k) 1_{\{b_j \neq a\}} + k c_j 1_{\{b_j = a\}} \right] b_j^{k-1}$$

Suppose that this last relation is true for a fixed $k \geq 0$. From Relation (7), we obtain

$$\begin{aligned} u_{k+1} &= \alpha a^{k+1} + \sum_{j=1}^{N} \left[a f_j(k) 1_{\{b_j \neq a\}} + k a c_j 1_{\{b_j = a\}} \right] b_j^{k-1} \\ &\quad + \sum_{j=1}^{N_j} (c_j b_j + k d_j) b_j^{k-1} \\ &= \alpha a^{k+1} + \sum_{j=1}^{N} \Big[a f_j(k) 1_{\{b_j \neq a\}} + k c_j b_j 1_{\{b_j = a\}} \\ &\qquad + c_j b_j + k d_j \Big] b_j^{k-1}. \end{aligned}$$

Writing $c_j b_j = c_j b_j 1_{\{b_j \neq a\}} + c_j b_j 1_{\{b_j = a\}}$ and $d_j = d_j 1_{\{b_j \neq a\}}$ since $d_j = 0$ when $b_j = a$, we obtain

$$\begin{aligned} u_{k+1} = \alpha a^{k+1} + \sum_{j=1}^{N} \Big[&(a f_j(k) + c_j b_j + k d_j) 1_{\{b_j \neq a\}} \\ &+ (k+1) c_j b_j 1_{\{b_j = a\}} \Big] b_j^{k-1}. \quad (9) \end{aligned}$$

The first term of this last summation can be simplified as follows. By definition of $f_j(k)$ and observing that $a\gamma_j + d_j = \gamma_j b_j$, we have

$$\begin{aligned} a f_j(k) + c_j b_j + k d_j &= a\theta_j b_j + c_j b_j + k(a\gamma_j + d_j) \\ &= a\theta_j b_j + c_j b_j + k\gamma_j b_j \\ &= (a\theta_j + c_j + k\gamma_j) b_j \\ &= (a\theta_j + c_j - \gamma_j + (k+1)\gamma_j) b_j. \end{aligned}$$

Since $c_j - \gamma_j = (b_j - a)\theta_j$, this last expression leads to

$$a f_j(k) + c_j b_j + k d_j = (\theta_j b_j + (k+1)\gamma_j) b_j = b_j f_j(k+1).$$

Putting this result into (9) gives

$$\begin{aligned} u_{k+1} = \alpha a^{k+1} + \sum_{j=1}^{N} \Big[&f_j(k+1) 1_{\{b_j \neq a\}} \\ &+ (k+1) c_j 1_{\{b_j = a\}} \Big] b_j^{k}, \end{aligned}$$

which completes the proof. ■

We are now ready to prove Theorem 3.

Proof of Theorem 3: The proof is made by recurrence on integer i. In fact, we prove that for every $i = 1, \ldots, n-1$, we have

$$V_{n-i}(k) = \sum_{j=1}^{n/2} (c_{i,j}(1 - p_j) + k d_{i,j}) (1 - p_j)^{k-1} \quad (10)$$

$$d_{i,j} = 0 \text{ for } j < n - i \quad (11)$$

$$c_{i,j} = 0 \text{ for } j > i. \quad (12)$$

Relations (11) and (12) are true for $i = 1$ since by definition we have $c_{1,j} = 1_{\{j=1\}}$ and $d_{1,j} = 0$. It follows that Relation (10) gives, for $i = 1$, $V_{n-1}(k) = (1 - p_1)^k$, which is in accordance with Relation (4).

Suppose now that Relations (10), (11) and (12) are true for a fixed integer i, $i \leq n - 2$. Using Relation (4) at point $k+1$ and the fact that $p_{n-i-1} = p_{i+1}$, we obtain

$$\begin{aligned} V_{n-i-1}(k+1) &= (1 - p_{i+1}) V_{n-i-1}(k) + p_{i+1} V_{n-i}(k) \\ &= (1 - p_{i+1}) V_{n-i-1}(k) \\ &\quad + p_{i+1} \sum_{j=1}^{n/2} (c_{i,j}(1 - p_j) + k d_{i,j}) (1 - p_j)^{k-1}. \end{aligned} \quad (13)$$

Integer i being fixed, we apply Lemma (9) in which we set $u_k = V_{n-i-1}(k)$, $a = 1 - p_{i+1}$, $N = \lfloor n/2 \rfloor$ and for $j = 1, \ldots, N$, $b_j = 1 - p_j$, $c_j = p_{i+1} c_{i,j}$ and $d_j = p_{i+1} d_{i,j}$. Note that condition of Lemma (9) is satisfied. Indeed, $a = b_j$ is equivalent to either $i + 1 = j$ or $i + 1 = n - j$. Since $i + 1 = j$ is equivalent to $i + j + 1 = 2j$ and since $j \leq n/2$, both conditions implies that $i + j < n$ which means from Relation (11) that $d_j = p_{i+1} d_{i,j} = 0$. With this notation, the parameters θ_j and γ_j writes

$$\begin{aligned} \theta_j &= \frac{c_j}{b_j - a} - \frac{d_j}{(b_j - a)^2} = \frac{p_{i+1} c_{i,j}}{p_{i+1} - p_j} - \frac{p_{i+1} d_{i,j}}{(p_{i+1} - p_j)^2}, \\ \gamma_j &= \frac{d_j}{b_j - a} = \frac{p_{i+1} d_{i,j}}{p_{i+1} - p_j}. \end{aligned}$$

Applying Lemma (9) from (13), we obtain

$$V_{n-i-1}(k) = \left(1 - \sum_{j=1}^{n/2} \theta_j 1_{\{i+1 \neq j, n-j\}}\right)(1-p_{i+1})^k$$
$$+ \sum_{j=1}^{n/2} (\theta_j(1-p_j) + k\gamma_j) 1_{\{i+1 \neq j, n-j\}} (1-p_j)^{k-1}$$
$$+ \sum_{j=1}^{n/2} k p_{i+1} c_{i,j} 1_{\{i+1=j \text{ or } i+1=n-j\}} (1-p_j)^{k-1}.$$

Defining

$$\begin{aligned}
c_{i+1,j} &= \theta_j && \text{for } i+1 \neq j, n-j,\\
d_{i+1,j} &= \gamma_j && \text{for } i+1 \neq j, n-j,\\
c_{i+1,i+1} &= 1 - \sum_{j=1}^{n/2} c_{i+1,j} && \text{for } i+1 \leq n/2,\\
c_{i+1,n-i-1} &= 1 - \sum_{j=1}^{n/2} c_{i+1,j} && \text{for } i+1 > n/2,\\
d_{i+1,i+1} &= p_{i+1} c_{i,i+1} && \text{for } i+1 \leq n/2,\\
d_{i+1,n-i-1} &= p_{i+1} c_{i,n-i-1} && \text{for } i+1 > n/2,
\end{aligned}$$

we obtain, using again the fact that $p_{i+1} = p_{n-i-1}$,

$$V_{n-i-1}(k) = c_{i+1,i+1} 1_{\{i+1 \leq n/2\}} (1-p_{i+1})^k$$
$$+ c_{i+1,n-i-1} 1_{\{n-i-1 < n/2\}} (1-p_{i+1})^k$$
$$+ \sum_{j=1}^{n/2} (c_{i+1,j}(1-p_j) + k d_{i+1,j}) 1_{\{i+1 \neq j, n-j\}} (1-p_j)^{k-1}$$
$$+ k p_{i+1} c_{i,i+1} 1_{\{i+1 \leq n/2\}} (1-p_{i+1})^{k-1}$$
$$+ k p_{i+1} c_{i,n-i-1} 1_{\{n-i-1 < n/2\}} (1-p_{i+1})^{k-1},$$

which gives

$$V_{n-i-1}(k) = \sum_{j=1}^{n/2} \left(c_{i+1,j}(1-p_j) + k d_{i+1,j}\right)(1-p_j)^{k-1}.$$

Now, we must prove the recurrence for the two additional properties (11) and (12). Note that $i+1 < j$ implies $i+1 \neq j$ and $i+1 \neq n-j$. So if $i+1 < j$ then we have $i < j$ and $i < n/2 < n-j$ which means that $c_{i,j} = d_{i,j} = 0$, which in turn implies that $\theta_j = 0$ and thus $c_{i+1,j} = 0$

In the same way, note that $i+1 < n-j$ implies $i+1 \neq n-j$. So if $i+1 < n-j$ then we have $i < n-j$ which means that $d_{i,j} = 0$, which in turn implies that $\gamma_j = 0$ and thus $d_{i+1,j} = 0$. This completes the proof. ■

Proof of Theorem 6: In order to simplify the writing, we introduce the following notations.

$$\begin{aligned}
A_n &= (n-1)H_{n-1} \ln\left(1 - \frac{2}{n}\right)\\
B_n &= \frac{2(n-1)H_{n-1}(n-2)^2}{n}\\
C_n &= \left(1 - \frac{2}{n}\right)^{(n-1)H_{n-1} - 2}.
\end{aligned}$$

Firstly, since

$$\begin{aligned}
\left(1 - \frac{2}{n}\right)^{\ln(c)(n-1)H_{n-1}} &= c^{(n-1)H_{n-1}\ln(1-2/n)}\\
&= c^{A_n}
\end{aligned}$$

we have

$$\frac{f(c,n)}{g(c,n)} = C_n(1 + cB_n)c^{A_n+1}.$$

Taking the derivative with respect to c, gives

$$\frac{\partial(f/g)}{\partial c}(c,n) = C_n c^{A_n}[A_n + 1 + cB_n(A_n+2)].$$

The term $C_n c^{A_n}$ is strictly positive for all $c \geq 1$, so we have

$$\frac{\partial(f/g)}{\partial c}(c,n) \geq 0 \Longleftrightarrow c \leq -\frac{A_n+1}{B_n(A_n+2)},$$

It is easy to check that $-(A_n+1)/(B_n(A_n+2)) < 0$ for all $n \geq 3$, which means that $\partial(f/g)/\partial c < 0$ for all $c \geq 1$. This implies that the function $c \longmapsto f(c,n)/g(c,n)$ is strictly decreasing on $[1, +\infty)$.

Observing that $A_n < -2$ for every $n \geq 3$, we obtain

$$\lim_{c \longrightarrow \infty} \frac{f(c,n)}{g(c,n)} = 0,$$

which proves the second part of the theorem. Secondly, since

$$\frac{f(1,n)}{g(1,n)} = C_n(1 + B_n),$$

it is easy to check that $f(1,n)/g(1,n) \geq 1$, for all $n \geq 3$.

Let us recapitulate and conclude. For all $n \geq 3$, we have $f(1,n)/g(1,n) \geq 1$ and $c \longmapsto f(c,n)/g(c,n)$ is continuous and strictly decreasing to 0. It follows that there exists a unique value $c* \geq 1$ such that $f(c^*,n)/g(c^*,n) = 1$ and satisfying the conditions (6). This completes the proof. ■