Randomized Rumor Spreading Revisited

Benjamin Doerr[1] and Anatolii Kostrygin[2]

1 **Laboratoire d'Informatique (LIX), École Polytechnique, Palaiseau, France**
`doerr@lix.polytechnique.fr`

2 **Laboratoire d'Informatique (LIX), École Polytechnique, Palaiseau, France**
`anatolii.kostrygin@gmail.com`

Abstract

We develop a simple and generic method to analyze randomized rumor spreading processes in fully connected networks. In contrast to all previous works, which heavily exploit the precise definition of the process under investigation, we only need to understand the probability and the covariance of the events that uninformed nodes become informed. This universality allows us to easily analyze the classic push, pull, and push-pull protocols both in their pure version and in several variations such as messages failing with constant probability or nodes calling a random number of others each round. Some dynamic models can be analyzed as well, e.g., when the network is a $G(n,p)$ random graph sampled independently each round [Clementi et al. (ESA 2013)].

Despite this generality, our method determines the expected rumor spreading time precisely apart from additive constants, which is more precise than almost all previous works. We also prove tail bounds showing that a deviation from the expectation by more than an additive number of r rounds occurs with probability at most $\exp(-\Omega(r))$.

We further use our method to discuss the common assumption that nodes can answer any number of incoming calls. We observe that the restriction that only one call can be answered leads to a significant increase of the runtime of the push-pull protocol. In particular, the double logarithmic end phase of the process now takes logarithmic time. This also increases the message complexity from the asymptotically optimal $\Theta(n \log\log n)$ [Karp, Shenker, Schindelhauer, Vöcking (FOCS 2000)] to $\Theta(n \log n)$. We propose a simple variation of the push-pull protocol that reverts back to the double logarithmic end phase and thus to the $\Theta(n \log\log n)$ message complexity.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems

Keywords and phrases Epidemic algorithm, rumor spreading, dynamic graph

Digital Object Identifier 10.4230/LIPIcs.ICALP.2017.138

1 Introduction

Randomized rumor spreading is one of the core primitives to disseminate information in distributed networks. It builds on the paradigm that nodes call random neighbors and exchange information with these contacts. This gives highly robust dissemination algorithms belonging to the broader class of gossip-based algorithms that, due to their epidemic nature, are surprisingly efficient and scalable. Randomized rumor spreading has found numerous applications, among others, maintaining the consistency of replicated databases [9], disseminating large amounts of data in a scalable manner [25], and organizing any kind of communication in highly dynamic and unreliable networks like wireless sensor networks and mobile ad-hoc networks [21]. Randomized rumor spreading processes are also

44th International Colloquium on Automata, Languages, and Programming (ICALP 2017).
Editors: Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl;
Article No. 138; pp. 138:1–138:14
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

used to model epidemic processes like viruses spreading over the internet [1], news spreading in social networks [10], or opinions forming in social networks [24].

The importance of these processes not only has led to a huge body of experimental results, but, starting with the influential works of Frieze and Grimmett [16] and Karp, Shenker, Schindelhauer, and Vöcking [22] also to a large number of mathematical analyses of rumor spreading algorithms giving runtime or robustness guarantees for existing algorithms and, based on such findings, proposing new algorithms.

Roughly speaking, two types of results can be found in the literature, general bounds trying to give a performance guarantee based only on certain graph parameters and analyses for specific graphs or graph classes. In the domain of general bounds, there is the classic maximum-degree-diameter bound of [13] and more recently, a number of works bounding the rumor spreading time in terms of conductance or other expansion properties [26, 6, 18, 19], which not only greatly helped our understanding of existing processes, but could also be exploited to design new dissemination algorithms [2, 3, 4, 20]. The natural downside of such general results is that they often do not give sharp bounds. It seems that among the known graph parameters, none captures very well how suitable this network structure is for randomized rumor spreading. Also, it has to be mentioned that these results mostly apply to the push-pull protocol.

The other research direction followed in the past is to try to prove sharper bounds for specific graph classes. This led, among others, to the results that the push-protocol spreads a rumor in a complete graph in time $\log_2 n + \ln n \pm \omega(1)$ with high probability $1 - o(1)$ (whp.) [28] (and in time $\log_{2-p} n + \frac{1}{p} \ln n \pm o(\log n)$ when messages fail independently with probability p), whereas the push-pull protocol does so in time $\log_3 n + O(\log \log n)$ [22]. The push protocol spreads rumors in hypercubes in time $O(\log n)$ whp. [13], determining the leading constant is a major open problem. For Erdős-Rény random graphs with edge probability asymptotically larger than the connectivity threshold, again a runtime of $\log_{2-p} n + \frac{1}{p} \ln n \pm o(\log n)$ was shown for the push protocol allowing transmission errors with rate p [14]. For preferential attachment graphs, which are often used as model for real-world networks, it was proven that the push-protocol needs $\Omega(n^\alpha)$ rounds, $\alpha > 0$ some constant, whereas the push-pull protocol takes time $\Theta(\log n)$ and $\Theta((\log n)/\log \log n)$ when nodes avoid to call the same neighbor twice in a row [5, 10]. Even faster rumor spreading times were shown on Chung-Lu power-law random graphs [15].

One weakness of all these results on specific graphs is that they very much rely on the particular properties of the protocol under investigation. Even in fully connected networks (complete graphs), the existing analyses for the basic push protocol [16, 28, 12], the push protocol in the presence of transmission failures [11], the push protocol with multiple calls [27], and the push-pull protocol [22] all uses highly specific arguments that cannot be used immediately for the other processes. This is despite the fact that the global behavior of these processes is often very similar. For example, all processes mentioned have an exponential expansion phase in which the number of informed node roughly grows by a constant factor until a constant fraction of the nodes is informed. Clearly, this hinders a faster development of the field. Note that the typical analysis of a rumor spreading protocol in the papers cited above needs between six and eight pages of proofs.

1.1 Our Results

In this work, we make a big step forward towards overcoming this weakness. We propose a *general analysis method* for all symmetric and memoryless rumor spreading processes in complete networks. It allows to easily analyze all rumor spreading processes mentioned above and many new ones. The key to this generality is showing that the rumor spreading times for these protocols are determined by the probabilities p_k of a new node becoming informed in a round starting with k informed nodes together with a mild bound on the covariance on the indicator random variables of the events that new nodes become informed. Consequently, all other particularities of the protocol can safely be ignored.

Despite this generality, our method gives bounds for the expected rumor spreading time that are *tight apart from an additive constant number of rounds*. Such tight bounds so far have only been obtained once, namely for the basic push protocol [12].

Our method also gives *tail bounds* stating that deviations from the expectation by an additive number of at least r of rounds occur with probability at most $A' \exp(-\alpha' r)$, where $A', \alpha' > 0$ are absolute constants. Such a precise tail bound was previously given only for the push protocol in [12]. Note that our tail bounds imply the usual whp-statements, e.g., that overshooting the expectation by any $\omega(1)$ term happens with probability $o(1)$ only, and that a rumor spreading time of $O(\log n)$ can be obtained with probability $1 - n^{-c}$, c any constant, by making the implicit constant in the time bound large enough.

We use our method to obtain the following particular results. We only state the expected runtimes. In all cases, the above tail bounds are valid as well.

Classic protocols, robustness: We start by analyzing the three basic push, pull, and push-pull protocols. In the *push protocol*, in each round each informed node calls a random node and sends a copy of the rumor to it. In the *pull protocol*, in each round each uninformed node calls a random node and tries to obtain the rumor from it. In the *push-pull protocol*, all nodes contact a random node and in each such contact the informed nodes send rumor to the communication partner.

For these three protocols, both in the fault-free setting and when assuming that calls fail independently with probability $1 - p$, our method easily yields the expected rumor spreading times given in Table 1. Note that all previous works apart from [12] did not state explicitly a bound for the expected runtime. Note further that for half of the settings regarded in Table 1 no previous result existed. In particular, we are the first to find that the double logarithmic shrinking phase observed by Karp et al. [22] for the push-pull protocol disappears when messages fail with constant probability p, and is instead replaced by an ordinary shrinking regime with the number of uninformed nodes reducing by roughly a factor of $(1-p)e^{-p}$ each round. This observation is not overly deep, but has the important consequence that the message complexity of the push-pull protocol raises from the theoretically optimal $\Theta(n \log\log n)$ value proven in [22] to an order of magnitude of $\Theta(n \log n)$ in the presence of a constant rate of transmission errors. Hence the significant superiority of the push-pull protocol over the push protocol in the fault-free setting reduces to a constant-factor advantage in the faulty setting.

Multiple calls: Panagiotou, Pourmiri, and Sauerwald [27] proposed a variation of the classic protocols in which the number of calls (always to different nodes) each node performs when active is a positive random variable R. They mostly assume that for each node, this random number is sampled once at the beginning of the process. For the case that R has constant expectation and variance, they show that the rumor spreading time of the push protocol is $\log_{1+\mathbb{E}[R]} n + \frac{1}{\mathbb{E}[R]} \ln n \pm o(\log n)$ with high probability and that the rumor spreading time of the push-pull protocol is $\Omega(\log n)$ with probability $1 - \varepsilon$, $\varepsilon > 0$. When R follows a power law with exponent $\beta = 3$, the push-pull protocol takes $\Theta(\frac{\log n}{\log\log n})$ rounds, and when $2 < \beta < 3$, it takes $\Theta(\log\log n)$ rounds.

Table 1 New and previous-best results for rumor spreading time T of the classic rumor spreading protocols in complete graphs on n vertices. The first line of each table entry contains the result that follows from the method proposed in this work, the second line states the best previous result (if any). For all new bounds on the expected rumor spreading time, a tail bound of type $\mathbb{P}[|T \geq \mathbb{E}[T]| \geq r] \leq A' \exp(-\alpha' r)$ with $A', \alpha' > 0$ suitable constants follows as well from this work. In [12], such a bound was given for the rumor spreading time of the push protocol without transmission failures.

	no transmission failures	calls fail indep. with prob. $1-p \in (0,1)$
push protocol	$\mathbb{E}[T] = \log_2 n + \ln n \pm O(1)$ $\lfloor \log_2 n \rfloor + \ln n - 1.116 \leq \mathbb{E}[T] \leq \lceil \log_2 n \rceil + \ln n + 2.765 + o(1)$ [12]	$\mathbb{E}[T] = \log_{1+p} n + \frac{1}{p} \ln n \pm O(1)$ $T = \log_{1+p} n + \frac{1}{p} \ln n \pm o(\log n)$ whp. [11]
pull protocol	$\mathbb{E}[T] = \log_2 n + \log_2 \ln n \pm O(1)$	$\mathbb{E}[T] = \log_{1+p} n + \frac{1}{\ln \frac{1}{1-p}} \ln n \pm O(1)$
push-pull protocol	$\mathbb{E}[T] = \log_3 n + \log_2 \ln n \pm O(1)$ $T = \log_3 n \pm O(\log \log n)$ whp. [22]	$\mathbb{E}[T] = \log_{1+2p} n + \frac{1}{p + \ln \frac{1}{1-p}} \ln n \pm O(1)$

The model of [27] makes sense when assuming that nodes have generally different communication capacities. To model momentarily different capacities, e.g., caused by being occupied with other communication tasks, we assume that the random variable is resampled for each node in each round. We also allow R to take the value 0. Again for the case $\mathbb{E}[R] = \Theta(1)$ and $\mathrm{Var}[R] = O(1)$, we show that the expected rumor spreading time of the push protocol is $\log_{1+\mathbb{E}[R]} n + \frac{1}{\mathbb{E}[R]} \ln n \pm O(1)$. The rumor spreading time of the push-pull protocol depends critically on the smallest value ℓ which R takes with positive probability. If $\ell = 0$, that is, with constant probability nodes contact no other node, then there is no double exponential shrinking and the expected rumor spreading time is $\log_{1+2\mathbb{E}[R]} n + \frac{1}{\mathbb{E}[R] - \ln \mathbb{P}[R=0]} \ln n \pm O(1)$. If nodes surely perform at least one call, then we have a double exponential shrinking regime and an expected rumor spreading time of $\log_{1+2\mathbb{E}[R]} n + \log_{1+\ell} \ln n \pm O(1)$.

Dynamic networks: We also show that our method is capable of analyzing dynamic networks when the dynamic is memory-less. Clementi et al. [7] have shown that when the network in each round is a newly sampled $G(n,p)$ random graph, then for any constant c the rumor spreading time of the push protocol is $\Theta(\log(n)/\min\{p, 1/n\})$ with probability $1 - n^{-c}$. We sharpen this result for the most interesting regime that $p = a/n$, a a positive constant. For this case, we show that the expected rumor spreading time is $\log_{2-e^{-a}} n + \frac{1}{1-e^{-a}} \ln(n) + O(1)$. Our tail bound $\mathbb{P}[|T - \mathbb{E}[T]| \geq r] \leq A' \exp(-\alpha' r)$ for suitable constants $A', \alpha > 0$ implies also the large deviation statement of [7] (where for $\Theta(\log n)$ deviations in the lower tail the trivial $\log_2(n)$ lower bound holding with probability 1 should be used).

Answering single calls only: We finally use our method to discuss an aspect mostly ignored by previous research. While in all protocols above (apart from the one of [27]) it is assumed that each node can call at most one other node per round, it is tacitly assumed in the pull and push-pull protocols that nodes can answer all incoming calls. For complete graphs on n vertices, the classic balls-into-bins theory immediately gives that in a typical round there is at least one node that receives $\Theta(\frac{\log n}{\log \log n})$ calls. So unlike for the outgoing traffic, nodes are implicitly assumed to be able to handle very different amounts of incoming traffic in one round.

The first to discuss this issue are Daum, Kuhn, and Maus [8] (also the SIROCCO 2016 best paper). Among other results, they show that if only one incoming call can be answered and if this choice is taken adversarially, then there are networks where a previously polylogarithmic rumor spreading time of the pull protocol becomes $\tilde{\Omega}(\sqrt{n})$. If the choice which incoming call is answered is taken randomly, then things improve and the authors show that for any network, the rumor spreading times of the pull and push-pull protocol increase by at most a factor of $O(\frac{\Delta(G)}{\delta(G)} \log n)$ compared to the variant in which all incoming calls are answered. Subsequently, Ghaffari and Newport [17] showed that with the restriction to accept only one incoming call, the general performance guarantees for the push-pull protocol in terms of vertex expansion or conductance [18, 19] do not hold. Kiwi and Caro [23] showed that solving the problem of multiple incoming calls via a FIFO queue can lead to extremely long rumor spreading times.

With our generic method, we can easily analyze this aspect of rumor spreading on complete graphs. While for the pull protocol only the growth phase mildly slows down, giving a total expected rumor spreading time of $\mathbb{E}[T] = \log_{2-1/e} n + \log_2 \ln n \pm O(1)$, for the push-pull protocol also the double logarithmic shrinking phase breaks down and we observe a total runtime of $\mathbb{E}[T] = \log_{3-2/e} n + \frac{1}{2} \ln n \pm O(1)$ and, similarly as for the push-pull protocol with transmission failures, an increase of the message complexity to $\Theta(n \log n)$. The reason, as our proof reveals, is that when a large number of nodes are informed, then their push calls have little positive effect (as in the classic push-pull protocol), but they now also block other nodes' pull calls from being accepted. This problem can be overcome by changing the protocol so that informed nodes stop calling others when the rumor is $\log_{3-2/e} n$ rounds old. The rumor spreading time of this modified push-pull protocol is $\mathbb{E}[T] = \log_{3-2/e} n + \log_2 \ln n \pm O(1)$ and, when halted at the right moment, this process takes $\Theta(n \log \log n)$ messages.

2 Outline of the Analysis Method

As just discussed, the main advantages of our approach are its universality and the very tight bounds it proves. We now briefly sketch the main new ideas that lead to this progress. Interestingly, they are rather simpler than the ones used in previous works.

2.1 Tight Bounds via a Target-Failure Calculus

We first describe how we obtain estimates for the rumor spreading time that are *tight apart from additive constants*. Let us take as example the classic push protocol. It is easy to compute that in a round starting with k informed nodes, the expected number of newly informed nodes is $E(k) = k - \Theta(k^2/n)$. Hence roughly speaking the number of informed nodes doubles each round (which explains the $\log_2 n$ part of the $\log_2 n + \ln n \pm O(1)$ rumor spreading time), but there is a growing gap to truly doubling caused by (i) calls reaching already informed nodes and (ii) several calls reaching the same target. This weakening of the doubling process was a main difficulty in all previous works.

The usual way to analyze this weakening doubling process is to partition the rumor spreading process in phases and within each phase to uniformly estimate the progress. For example, Pittel [28] considers 7 phases. He argues first that with high probability the number if informed nodes doubles until $n_1 = o(\sqrt{n})$ nodes are informed. Then, until $n_2 = n/\log^2(n)$ nodes are informed, with high probability in each round the number of informed nodes increases by at least a factor of $2(1 - \frac{1}{\log^2(n)})$. Consequently, this second phase lasts at most $\log_{2(1-\frac{1}{\log^2(n)})}(n_2/n_1)$ rounds. While this type of argument gives good bounds for phases bounded away from the middle regime with both $\Theta(n)$ nodes informed and uninformed, we do not see how this "estimating a phase uniformly" argument can cross the middle regime without losing a number $\omega(1)$ of rounds.

For this reason, we proceed differently. To prove upper bounds on rumor spreading times, for each number k of informed nodes, we formulate a pessimistic round target $E_0(k)$ that is sufficiently below the expected number $E(k)$ of newly informed nodes. Here "sufficiently below" means that the probability $q(k)$ to fail reaching this target number of informed nodes is small, but not necessarily $o(1)$ as in all previous analyses. Using a restart argument, we observe that the random time needed to go from k informed nodes to at least $E_0(k)$ informed nodes is stochastically dominated by 1 plus a geometric random variable with parameter $1-q(k)$, where all our geometric random variables count the number of failures until success (this is one of the two definitions of geometric distributions that are in use). In particular, the expected time to go from k to at least $E_0(k)$ informed nodes is at most $1+\frac{q(k)}{1-q(k)}$.

The second, again elementary, key argument is that when we define a sequence of round targets by $k_0 := 1$, $k_1 := E_0(k_0)$, $k_2 := E_0(k_1), \dots$ with suitably defined $E_0(\cdot)$, then the k_i grow almost like 2^i (in the example of the classic push protocol). More precisely, there is a $T = \log_2 n \pm O(1)$ such that $k_T = \Theta(n)$. Hence together with the previous paragraph we obtain that the number of rounds to reach k_T informed nodes is dominated by T plus a sum of independent geometric random variables. This sum has expectation $\sum_{i=0}^{T-1} \frac{q(k_i)}{1-q(k_i)} = O(\sum_{i=0}^{T-1} q(k_i))$, so it suffices that the sum of the failure probabilities $q(k_i)$ is a constant (unlike in previous works, where it needed to be $o(1)$). A closer look at this sum also gives the desired tail bounds.

Similarly, to prove matching lower bounds, we define optimistic round targets $E_0(k)$ such that a round starting with k informed nodes finds it unlikely to reach $E_0(k)$ informed nodes. Since again we want to allow failure probabilities that are constant, we now have to be more careful and also quantify the probability to overshoot $E_0(k)$ by larger quantities. This will then allow to argue that when defining a sequence of round targets recursively as above, then the expected number of targets overjumped (and thus the expected number of rounds saved compared to the "one target per round" calculus), is only constant.

We remark that a target-failure argument similar to ours was used already in [12], there however only to give an upper bound for the runtime of the push protocol in the regime from n^s, s a small constant, to $\Theta(n)$ informed nodes, that is, the later part of the exponential growth regime of the push process, in which via Chernoff bounds very strong concentration results could be exploited. Hence the novelty of this work with respect to the target-failure argument is that this analysis method can be used (i) also from the very beginning of the process on, where we have no strong concentration, (ii) also for the exponential and double exponential shrinking regimes of rumor spreading processes, and (iii) also for lower bounds.

2.2 Uniform Treatment of Many Rumor Spreading Processes

As discussed earlier, the previous works regarding different rumor spreading processes on complete graphs all had to use different arguments. The reason is that the processes, even when looking similar from the outside, are intrinsically different when looking at the details. As an example, let us consider the first few rounds of the push and the pull protocol. In the push protocol, we just saw that while there are at most $o(\sqrt{n})$ nodes informed, then a birthday paradox type argument gives that with high probability we have perfect doubling in each round. For the pull process, in which each uninformed node calls a random node and becomes informed when the latter was informed, we also easily compute that a round starting with k informed nodes creates an expected number of $(n-k)\frac{k}{n} = k - \frac{k^2}{n}$ newly informed nodes. However, since these are binomially distributed, there is no hope for perfect doubling. In fact, for the first constant number of rounds, we even have a constant probability that not a single node becomes informed.

The only way to uniformly treat such different processes is by making the analysis depend only on general parameters of the process as opposed to the precise definition. Our second main contribution is distilling a few simple conditions that (i) subsume essentially all symmetric and time-invariant rumor spreading processes on complete graphs and (ii) suffice to prove rumor spreading times via the above described target-failure method. All this is made possible by the observation that the target-failure method needs much less in terms of failure probabilities than previous approaches, in particular, it can tolerate constant failure probabilities. Consequently, instead of using Chernoff and Azuma bounds for independent or negatively correlated random variables (which rely on the precise definition of the process), it suffices to use Chebyshev's inequality as concentration result.

Consequently, to apply our method we only need to (i) understand (with a certain precision) the probability p_k that an uninformed node becomes informed in a round starting with k informed nodes; recall that we assumed symmetry, that is, this probability is the same for all uninformed nodes, and (ii) we need to have a mild upper bound on the covariance of the indicator random variables of the events that two nodes become informed.

The probabilities p_k usually are easy to compute from the protocol definition. Also, we do not know them precisely. For example, for the growth phase of the push protocol discussed above, it suffices to know that there are constants $a < 2$ and a' such that for all $k < n/2$ we have $\frac{k}{n}(1 - a\frac{k}{n}) \le p_k \le \frac{k}{n}(1 + a'\frac{k}{n})$. This (together with the covariance condition) is enough to show that the rumor spreading process takes $\log_2 n \pm O(1)$ rounds to inform $n/2$ nodes or more. The constants a, a' have no influence on the final result apart from the additive constant number of rounds hidden in the $O(1)$ term. The covariances are also often easy to bound with sufficient precision, among others, because many in processes the events that two uniformed nodes become informed are independent or negatively correlated.

In our general analysis method, we profit from the fact that seemingly all reasonable rumor spreading processes in complete networks can be described via three regimes:

Exponential growth: Up to a constant fraction fn of informed nodes, $p_k = \gamma_n \frac{k}{n}(1 \pm O(\frac{k}{n}))$. The number of informed nodes thus increases roughly by a factor of $(1 + \gamma_n)$ in each round, hence the expected time to reach fn informed nodes or more is $\log_{1+\gamma_n} n \pm O(1)$.

Exponential shrinking: From a certain constant fraction $u = n - k = gn$ of uninformed nodes on, the probability of remaining uninformed satisfies $1 - p_{n-k} = e^{-\rho_n} \pm O(\frac{u}{n})$. This leads to a shrinking of the number of uninformed nodes by essentially a factor of $e^{-\rho_n}$ per round. Hence when starting with gn informed nodes, it takes another $\frac{1}{\rho_n} \ln n \pm O(1)$ rounds in expectation until all are informed.

Double exponential shrinking: From a certain constant fraction $u = n - k = gn$ of uninformed nodes on, the probability of remaining uninformed satisfies $1 - p_{n-k} = \Theta((\frac{u}{n})^{\ell-1})$. Now the expected time to go from gn uninformed nodes to no uninformed node is $\log_\ell \ln n \pm O(1)$.

Due to their different nature, we cannot help treating these three regimes separately, however all with the target-failure method. Hence the main differences between these regimes lie in defining the pessimistic estimates for the targets, computing the failure probabilities, and computing the number of intermediate targets until the goal is reached. All this only needs computing expectations, using Chebyshev's inequality, and a couple of elementary estimates.

3 Precise Statement of the Technical Results

In this work, we consider only *homogeneous rumor spreading processes* characterized as follows. We always assume that we have n nodes. Each node can be either *informed* or *uninformed*. We assume that the process starts with exactly one node being informed. Uninformed nodes may become informed, but an informed node never becomes uninformed. We consider a discrete time process, so the process can be partitioned into *rounds*. In each round each uninformed node can become informed. Whenever a round starts with k nodes being informed, then the probability for each uninformed node to become informed is some number p_k, which only depends on the number k of informed nodes at the beginning of the round.

The main insight of this work is that for such homogeneous rumor spreading processes we can mostly ignore the particular structure of the process and only work with the *success probabilities* p_k defined above and the *covariance numbers* c_k defined as follows.

▶ **Definition 1** (Covariance numbers). For a given homogeneous rumor spreading process and $k \in [1..n-1]$ let c_k be the smallest number such that whenever a round starts with k informed nodes and for any two uninformed nodes x_1, x_2, the indicator random variables X_1, X_2 for the events that these nodes become informed in this round satisfy

$$\mathrm{Cov}[X_1, X_2] \le c_k.$$

Upper bound for these covariances imply upper bounds on the variance of the number of nodes newly informed in a round. If the latter is small, Chebyshev's inequality yields that the actual number of newly informed nodes deviates not a lot from its expectation (which is determined by p_k).

Our main interest is studying after how many round all nodes are informed.

▶ **Definition 2** (Rumor spreading times). Consider a homogeneous rumor spreading process. For all $t = 0, 1, \ldots$ denote by I_t the number of informed nodes at the end of the t-th round ($I_0 := 1$). Let $k \le m \le n$. By $T(k, m)$ we denote the time it takes to increase the number of informed nodes from k to m or more, that is,

$$T(k,m) = \min\{t - s | I_s = k \, \text{and} \, I_t \ge m\}.$$

We call $T(1, n)$ the rumor spreading time of the process.

As it turns out, almost all homogeneous rumor spreading processes can be analyzed via three regimes.

3.1 Exponential Growth Regime

When not too many nodes are informed, in most rumor spreading processes we observe roughly a constant-factor increase of the number of informed nodes in one round, however, this increase becomes weaker with increasing number of informed nodes.

▶ **Definition 3** (Exponential growth conditions). Let γ_n be bounded between two positive constants. Let $a, b, c \ge 0$ and $0 < f < 1$. We say that a homogeneous rumor spreading process satisfies the *upper (respectively lower) exponential growth conditions* in $[1, fn[$ if for any $n \in \mathbb{N}$ big enough the following properties are satisfied for any $k < fn$.

(i) $p_k \ge \gamma_n \frac{k}{n} \cdot \left(1 - a\frac{k}{n} - \frac{b}{\ln n}\right)$ (respectively $p_k \le \gamma_n \frac{k}{n} \cdot \left(1 + a\frac{k}{n} + \frac{b}{\ln n}\right)$).

(ii) $c_k \le c\frac{k}{n^2}$.

In the case of the upper exponential growth condition, we also require $af < 1$.

These growth conditions suffice to prove that in an expected time of at most (respectively at least) $\log_{1+\gamma_n} n \pm O(1)$ rounds a linear number of nodes becomes informed. Consequently, the decrease of the dissemination speed when more nodes are informed (quantified by the term $-a\frac{k}{n}$ in the upper exponential growth condition), which was a main difficulty in previous analyses, has only an $O(1)$ influence on the rumor spreading time.

▶ **Theorem 4.** *If a homogeneous rumor spreading process satisfies the upper (lower) exponential growth conditions in* $[1, fn[$, *then there are constants* $A', \alpha' > 0$ *such that*

$$\mathbb{E}[T(1, fn)] \underset{(\geq)}{\leq} \log_{1+\gamma_n} n \underset{(-)}{+} O(1),$$
$$\mathbb{P}[T(1, fn) \underset{(\leq)}{\geq} \log_{1+\gamma_n} n \underset{(-)}{+} r] \leq A' \exp(-\alpha' r) \text{ for all } r \in \mathbb{N}.$$

When the lower exponential growth conditions are satisfied, then also there is an $f' \in]f, 1[$ *such that with probability* $1 - O\left(\frac{1}{n}\right)$ *at most* $f'n$ *nodes are informed at the end of round* $T(1, fn)$.

We note that the upper tail bound is tight apart from the implicit constants. This is witnessed, for example, by the pull protocol, where rounds starting with only a constant number of informed nodes have a constant probability of not informing any new node.

3.2 Exponential Shrinking Regime

In a sense dual to the previous regime, in many rumor spreading processes we observe that the number of uninformed nodes shrinks by a constant factor once sufficiently many nodes are informed. Again, the weaker shrinking at the beginning of this regime has only an $O(1)$ influence on the resulting rumor spreading times.

▶ **Definition 5** (Exponential shrinking conditions)**.** Let ρ_n be bounded between two positive constants. Let $0 < g < 1$, and $a, c \in \mathbb{R}_{\geq 0}$. We say that a homogeneous rumor spreading process satisfies the *upper (respectively lower) exponential shrinking conditions* if for any $n \in \mathbb{N}$ big enough, the following properties are satisfied for all $u = n - k \leq gn$.

(i) $1 - p_k = 1 - p_{n-u} \leq e^{-\rho_n} + a\frac{u}{n}$ (respectively $1 - p_k = 1 - p_{n-u} \geq e^{-\rho_n} - a\frac{u}{n}$).

(ii) $c_k = c_{n-u} \leq \frac{c}{u}$.

For the upper exponential shrinking conditions, we also assume that $e^{-\rho_n} + ag < 1$.

▶ **Theorem 6.** *If a homogeneous rumor spreading process satisfies the upper (lower) exponential shrinking conditions, then there are* $A'\alpha' > 0$ *such that*

$$\mathbb{E}[T(n - \lfloor gn \rfloor, n)] \underset{(\geq)}{\leq} \frac{1}{\rho_n} \ln n \underset{(-)}{+} O(1),$$
$$\mathbb{P}[T(n - \lfloor gn \rfloor, n) \underset{(\leq)}{\geq} \frac{1}{\rho_n} \ln n \underset{(-)}{+} r] \leq A' \exp(-\alpha' r) \text{ for all } r \in \mathbb{N}.$$

Again, the upper tail bound is tight apart from the constants as shown by the last rounds of the push protocol.

3.3 Double Exponential Shrinking Regime

Protocols using pull operations in the absence of transmission failures display a faster reduction of the number if uninformed nodes.

▶ **Definition 7** (Double exponential shrinking conditions)**.** Let $g \in]0,1]$, $\ell > 1$, and $a, c \in \mathbb{R}_{\geq 0}$ such that $ag^{\ell-1} < 1$. We say that a homogeneous rumor spreading process satisfies the *upper (respectively lower) double exponential shrinking conditions* if for any n big enough the following properties are satisfied for all $u = n - k \in [1, gn]$.

(i) $1 - p_{n-u} \leq a\left(\frac{u}{n}\right)^{\ell-1}$ (respectively $1 - p_{n-u} \geq a\left(\frac{u}{n}\right)^{\ell-1}$).

(ii) $c_{n-u} \leq c\frac{n}{u^2}$.

▶ **Theorem 8.** *If a homogeneous rumor spreading process satisfies the upper (lower) double exponential shrinking conditions, then there are $A', \alpha' > 0$ and R (depending on α) such that*

$$\mathbb{E}[T(n - \lfloor gn \rfloor, n)] \underset{(\geq)}{\leq} \log_\ell \ln n \underset{(-)}{+} O(1),$$

$$\mathbb{P}[T(n - \lfloor gn \rfloor, n) \geq \log_\ell \ln n + r] \leq O(n^{-\alpha' r + A'}) \text{ for all } r \in \mathbb{N},$$

$$(\mathbb{P}[T(n - \lfloor gn \rfloor, n) \leq \log_\ell \ln n - R] \leq O(n^{-1+2\ell\alpha})).$$

The last rounds of the push-pull protocol show that the upper tail bound is tight apart from the constants. The lower tail bound is clearly not best possible, but most likely good enough for most purposes.

3.4 Connecting Regimes

While often these above described three regimes suffice to fully analyze a rumor spreading process, occasionally it is necessary or convenient to separately regard a constant number of rounds between the growth and the shrinking regime. This is achieved by the following two lemmas.

▶ **Lemma 9.** *Consider a homogeneous rumor spreading process. Let $0 < \ell < m < n$ and $0 < p < 1$. Suppose for any number $\ell \leq k < m$, we have $p_k \geq p$. Then*

$$\mathbb{E}[T(\ell, m)] \leq \tfrac{n-\ell}{n-m} \cdot \tfrac{1}{p},$$

$$\mathbb{P}[T(\ell, m) > r] \leq \tfrac{n-\ell}{n-m} \cdot (1-p)^r \text{ for all } r \in \mathbb{N}.$$

▶ **Lemma 10.** *Let $f, p \in]0,1[$ and $c > 0$. Suppose that for any $k < fn$ we have $p_k \leq p$ and $c_k \leq \frac{c}{n}$. Then there exists $f' \in]f, 1[$ such that with probability $1 - O\left(\frac{1}{n}\right)$ at the end of some round the number of informed nodes will be between fn and $f'n$.*

4 Applying the Above Technical Results

In this section, we sketch how to use the above tools to obtain some of the results described in Section 1.1. To ease the notation we always assume that nodes call random nodes, that is, including themselves. The main observation is that computing the p_k is usually very elementary. For the covariance conditions, often we easily observe a negative or zero covariance, but when this is not true, then things can become technical.

For the *basic push, pull, and push-pull protocols*, we easily observe that all covariances to be regarded are negative or zero: Knowing that one uninformed node x_1 becomes informed in the current round has no influence on the pull call of another uninformed node x_2. When the protocol has push calls and x_1 was informed via a push call, then this event makes it slightly less likely that x_2 becomes informed via a push call, simply because at least one informed node is occupied with calling x_1.

The success probabilities p_k are easy to compute right from the protocol definition. When k nodes are informed, then the probabilities that an uninformed node becomes informed are

$$p_k = \begin{cases} 1-(1-1/n)^k & \text{for the push protocol,} \\ k/n & \text{for the pull protocol,} \\ p_k = 1-(1-1/n)^k \frac{n-k}{n} & \text{for the push-pull protocol.} \end{cases}$$

Using elementary estimates like $1-k/n \le (1-1/n)^k \le 1-k/n+k^2/2n^2$, we see that the push and pull protocols satisfy the exponential growth conditions with $\gamma_n = 1$, whereas the push-pull protocol does the same with $\gamma_n = 2$. The push protocol satisfies the exponential shrinking conditions with $\rho_n = 1$. The pull and push-pull protocols satisfy the double exponential shrinking conditions with $\ell = 2$. All growth conditions are satisfied at least up to $k = n/2$ informed nodes and all shrinking conditions are satisfied at least for $u \le n/2$ uninformed nodes, so we do not need the intermediate lemmas. This proves our results given in Table 1 for the fault-free case.

Faulty communication: The same arguments (with different constants γ_n and ρ_n) suffice to analyze these protocols when messages get lost independently with probability $1-p$. The only structural difference is that now for the pull and push-pull protocols uninformed nodes remain uninformed with at least constant probability. For this reason, now all three protocols have an exponential shrinking phase.

The push-pull protocol with the restriction that nodes *answer only a single incoming call* randomly chosen among the incoming calls is an example where the exponential growth and shrinking conditions are harder to prove. To compute the p_k we assume that all n calls have a random unique priority in $[1..n]$ and that the call with lowest priority number is accepted. For fixed priority, the probability of being accepted is easy to compute, and this leads to the success probability of a pull call. For the probability to become informed via a push call, the simple argument that the first incoming call is from an informed node with probability k/n solves the problem. When showing the covariance conditions, we face the problem that it is indeed not clear if we have negative or zero covariance. The event that some node becomes informed increases the chance that this node received a push call. This push call cannot interfere with another node's pull call to an informed node. So it does have some positive influence on the probability of another uninformed node to become informed. Fortunately, for our covariance conditions allow some positive correlation. Because of this, very generally speaking, we can ignore certain difficulties to handle situations when they occur rare enough.

Dynamic communication graphs: Being maybe the result where it is most surprising that bounds sharp apart from additive constants can be obtained, we now regard in more detail a problem regarded in [7]. There, the performance of push rumor spreading in a group of n agents was investigated when the actual communication network is changing in each round. As one such dynamic models, it was assumed that the communication graph in each round is a newly sampled $G(n, p)$ random graph, that is, there is an edge independently with probability p between any two vertices. For the ease of presentation, we assume that the edge probability equals $p = a/n$ for some constant $a > 0$. This is clearly the most interesting case. For such (and larger) p, a rumor spreading time of $\Theta(\log n)$ was shown to hold with inverse-polynomial failure probability. Recalling that for $p = a/n$ the graph $G(n, p)$ is not connected and has vertex degrees ranging from 0 to $\Theta(\log(n)/\log\log(n))$, this result is not obvious. Also, observe that the actions of all nodes in one round take place in the same random graph, so there are dependencies that have to be taken into account.

▶ **Theorem 11.** *Let T be the time the push protocol needs to inform n nodes when in each round a newly sampled $G(n,p)$, $p = a/n$, random graph represents the communication network. Then*

$$\mathbb{E}[T] = \log_{2-e^{-a}} n + \tfrac{1}{1-e^{-a}} \ln(n) \pm O(1)$$

and there are constants $A', \alpha' > 0$ such that $\mathbb{P}[|T - \mathbb{E}[T]| \geq r] \leq A' \exp(\alpha' r)$ holds for all $r \in \mathbb{N}$.

Recall that a vertex is isolated with probability $e^{-a} + O(1/n)$. Clearly, an informed vertex when isolated necessarily fails to inform another vertex in this round. The rumor spreading time proven above is the same as the one for the case that the communication network is always a complete graph, but calls fail independently with probability e^{-a}. Hence in a sense the changing topology (with low vertex degrees) is not harmful apart from the effect that it creates isolated vertices with constant rate. We did not expect this.

To prove Theorem 11, we first observe that the covariance properties are fulfilled. By symmetry, we can assume that in a round starting with k informed nodes, we first sample the random graph and decide for each node which neighbor it potentially calls in this round, and only then decide randomly which k nodes are informed and have these call the random neighbor determined before. Conditioning on the outcome of random graph, neighbor choice, and on that nodes x and y are not informed, in the remaining random experiment the events "x becomes informed" and "y becomes informed" clearly are negatively correlated.

Estimating the probability p_k for an uninformed node to become informed in a round starting with k informed nodes, is slightly technical. Since it is unlikely that two neighbors of an uninformed node x are connected by an edge, the main contribution to p_k stems from the case that the informed neighbors of x form an independent set. Conditioning on this outcome of the edges in $\{x\} \cup N(x)$, each informed neighbor of x has an independent probability of roughly $(1 - e^{-a})/a$ of calling x, giving (again taking care of the dependencies) a probability of roughly $1 - p_k \approx (1 - \frac{a}{n}\frac{1-e^{-a}}{a})^k$ for the event that no informed node calls x. From this, we estimate $\frac{k}{n}(1 - e^{-a})(1 - \frac{k+O(1)}{2n}(1 - e^{-a})) \leq p_k \leq \frac{k}{n}(1 - e^{-a} + O(1/n))$, showing that the exponential growth conditions are satisfied with $\gamma_n = 1 - e^{-a}$. Similar arguments, again taking some care for the dependencies that the random graph imposes on the actions of informed neighbors, show that the upper exponential shrinking conditions are satisfied for $\rho_n = 1 - e^{-a}$, whereas the lower exponential shrinking conditions are satisfied with $\rho_n = 1 - e^{-a} + O(\log(n)^2/n)$.

5 Summary, Outlook

In this work, we presented a general, easy-to-use method to analyze homogeneous rumor spreading processes on complete networks (including memoryless dynamic settings). Such processes are important in many applications, among others, due to the use of random peer sampling services in many distributed systems. Such processes also correspond to the fully mixed population model in mathematical epidemiology.

The two main strengths of our method are (i) that it builds only on estimates for the probability and the covariance of the events that new nodes become informed—consequently, many processes can be analyzed with identical arguments (as opposed to all previous works), and (ii) that it determines the expected rumor spreading time precise apart from additive constants (with tail bounds giving in most cases that deviations by an additive number r of rounds occur with probability $\exp(-\Omega(r))$ only). The key to our results is distilling growth and shrinking conditions which cover essentially all previously regarded homogeneous processes and which imply the desired runtime bounds.

From a broader perspective, this work shows that the traditional approach to randomized processes of splitting the analysis in several phases and then trying to understand each phase with uniform arguments might not be the ideal way to capture the nature of processes with a behavior changing continuously over time. While we demonstrated that the more careful round-target approach is better suited for homogeneous rumor spreading processes, one can speculate if similar ideas are profitable for other randomized algorithms or processes regarded in computer science.

References

1 Noam Berger, Christian Borgs, Jennifer T. Chayes, and Amin Saberi. On the spread of viruses on the internet. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 301–310. SIAM, 2005. `doi:10.1145/1070432.1070475`.

2 Keren Censor-Hillel, Bernhard Haeupler, Jonathan A. Kelner, and Petar Maymounkov. Global computation in a poorly connected world: fast rumor spreading with no dependence on conductance. In *Proceedings of the 44th Symposium on Theory of Computing Conference (STOC)*, pages 961–970, 2012. `doi:10.1145/2213977.2214064`.

3 Keren Censor-Hillel and Hadas Shachnai. Partial information spreading with application to distributed maximum coverage. In *Proceedings of the 29th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 161–170, 2010. `doi:10.1145/1835698.1835739`.

4 Keren Censor-Hillel and Hadas Shachnai. Fast information spreading in graphs with large weak conductance. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 440–448, 2011. `doi:10.1137/1.9781611973082.35`.

5 Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. Rumor spreading in social networks. In *Proceedings of the 36th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 375–386. Springer, 2009. `doi:10.1007/978-3-642-02930-1_31`.

6 Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. Almost tight bounds for rumour spreading with conductance. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 399–408. ACM, 2010. `doi:10.1145/1806689.1806745`.

7 Andrea E. F. Clementi, Pierluigi Crescenzi, Carola Doerr, Pierre Fraigniaud, Francesco Pasquale, and Riccardo Silvestri. Rumor spreading in random evolving graphs. *Random Structures and Algorithms*, 48:290–312, 2016. `doi:10.1002/rsa.20586`.

8 Sebastian Daum, Fabian Kuhn, and Yannic Maus. Rumor spreading with bounded in-degree. *CoRR*, abs/1506.00828, 2015.

9 Alan J. Demers, Daniel H. Greene, Carl Hauser, Wes Irish, John Larson, Scott Shenker, Howard E. Sturgis, Daniel C. Swinehart, and Douglas B. Terry. Epidemic algorithms for replicated database maintenance. In *Proceedings of the Sixth Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 1–12. ACM, 1987. `doi:10.1145/41840.41841`.

10 Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. Social networks spread rumors in sublogarithmic time. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 21–30. ACM, 2011. `doi:10.1145/1993636.1993640`.

11 Benjamin Doerr, Anna Huber, and Ariel Levavi. Strong robustness of randomized rumor spreading protocols. *Discrete Applied Mathematics*, 161:778–793, 2013. `doi:10.1016/j.dam.2012.10.014`.

12 Benjamin Doerr and Marvin Künnemann. Tight analysis of randomized rumor spreading in complete graphs. In *Proceedings of the Eleventh Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 82–91. SIAM, 2014. doi:10.1137/1.9781611973204.8.

13 Uriel Feige, David Peleg, Prabhakar Raghavan, and Eli Upfal. Randomized broadcast in networks. *Random Structures and Algorithms*, 1:447–460, 1990. doi:10.1002/rsa.3240010406.

14 Nikolaos Fountoulakis, Anna Huber, and Konstantinos Panagiotou. Reliable broadcasting in random networks and the effect of density. In *Proceedings of the 29th International Conference on Computer Communications (INFOCOM)*, pages 2552–2560. IEEE, 2010. doi:10.1109/INFCOM.2010.5462084.

15 Nikolaos Fountoulakis, Konstantinos Panagiotou, and Thomas Sauerwald. Ultra-fast rumor spreading in social networks. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1642–1660. SIAM, 2012.

16 Alan M. Frieze and Geoffrey R. Grimmett. The shortest-path problem for graphs with random arc-lengths. *Discrete Applied Mathematics*, 10:57–77, 1985. doi:10.1016/0166-218X(85)90059-9.

17 Mohsen Ghaffari and Calvin Newport. How to discreetly spread a rumor in a crowd. *CoRR*, abs/1607.05697, 2016.

18 George Giakkoupis. Tight bounds for rumor spreading in graphs of a given conductance. In *28th International Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 57–68. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2011. doi:10.4230/LIPIcs.STACS.2011.57.

19 George Giakkoupis. Tight bounds for rumor spreading with vertex expansion. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 801–815. SIAM, 2014. doi:10.1137/1.9781611973402.59.

20 Bernhard Haeupler. Simple, fast and deterministic gossip and rumor spreading. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 705–716, 2013. doi:10.1137/1.9781611973105.51.

21 Konrad Iwanicki and Maarten van Steen. Gossip-based self-management of a recursive area hierarchy for large wireless sensornets. *IEEE Transactions on Parallel and Distributed Systems*, 21:562–576, 2010. doi:10.1109/TPDS.2009.89.

22 Richard M. Karp, Christian Schindelhauer, Scott Shenker, and Berthold Vöcking. Randomized rumor spreading. In *Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS)*, pages 565–574. IEEE, 2000.

23 Marcos A. Kiwi and Christopher Thraves Caro. FIFO queues are bad for rumor spreading. *IEEE Trans. Information Theory*, 63:1159–1166, 2017. doi:10.1109/TIT.2016.2632153.

24 Jon M. Kleinberg. The convergence of social and technological networks. *Communications of the ACM*, 51:66–72, 2008. doi:10.1145/1400214.1400232.

25 Miguel Matos, Valerio Schiavoni, Pascal Felber, Rui Oliveira, and Etienne Riviere. BRISA: combining efficiency and reliability in epidemic data dissemination. In *26th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 983–994, 2012. doi:10.1109/IPDPS.2012.92.

26 Damon Mosk-Aoyama and Devavrat Shah. Fast distributed algorithms for computing separable functions. *IEEE Trans. Information Theory*, 54:2997–3007, 2008. doi:10.1109/TIT.2008.924648.

27 Konstantinos Panagiotou, Ali Pourmiri, and Thomas Sauerwald. Faster rumor spreading with multiple calls. *The Electronic Journal of Combinatorics*, 22:P1.23, 2015.

28 Boris Pittel. On spreading a rumor. *SIAM Journal on Applied Mathematics*, 47:213–223, 1987.