Tight Analysis of Randomized Rumor Spreading in Complete Graphs

Benjamin Doerr∗†‡ Marvin Künnemann†‡

Abstract

We present a tight analysis of the basic randomized rumor spreading process in complete graphs introduced by Frieze and Grimmett (1985), where in each round of the process each node knowing the rumor gossips the rumor to a node chosen uniformly at random. The process starts with a single node knowing the rumor.

We show that the number $S_n$ of rounds required to spread a rumor in a complete graph with $n$ nodes is very closely described by $\log_2 n$ plus $(1/n)$ times the completion time of the coupon collector process. This in particular gives very precise bounds for the expected runtime of the process, namely $\lceil \log_2 n \rceil + \ln n - 1.116 \leq \mathbb{E}[S_n] \leq \lceil \log_2 n \rceil + \ln n + 2.765 + o(1)$.

1 Introduction

Randomized rumor spreading are a class of randomized processes with ample applications in algorithmics, but also in modeling natural or technical spreading processes (epidemics, computer viruses). In this work, we shall give a very precise analysis of the most basic rumor spreading process in which a single piece of information is spread in a group of $n$ people by, in a round-based fashion, each informed person calling a random one and gossiping the rumor to him/her.

1.1 Randomized Rumor Spreading Processes

A randomized rumor spreading process is characterized by the fact that a rumor is spread in a network by nodes of the network exchanging information with randomly chosen neighbors. Such processes and similar ones have been studied in mathematical epidemiology and stochastic particle systems, see, e.g., [Lig99]. In computer science, besides modeling epidemic processes with relevance to computer science (spread of information in social networks [DFF12], spread of computer viruses [BBCS05], forming of opinions in social networks [Kle08]), rumor spreading is an important algorithmic paradigm.

While its very first occurrence [FG85] was only as an analysis tool for algorithmic problems with no immediate connection to rumor spreading, it quickly was noted that randomized rumor spreading can be used as a highly scalable and robust mechanism to distribute updates in replicated database applications [DGH+87]. This scalability today is mostly exploited in data-intensive applications, e.g., for media content or news feeds [MSF+12]. The second main application area of rumor

∗LIX, École Polytechnique, Palaiseau, France
†Max-Planck-Institut für Informatik, Saarbrücken, Germany
‡Universität des Saarlandes, Saarbrücken, Germany
spreading (here often called gossip-based algorithms) are wireless sensor networks and mobile ad-hoc networks. Here, the simple paradigm of contacting random neighbors is used to overcome the difficulties imposed by the changing and unreliable network topology, see, e.g., [IvS10].

1.2 Basic Rumor Spreading

Possibly the most basic rumor spreading process regarded already in the paper by Frieze and Grimmett [FG85] is the following, aiming to spread a rumor in a network of \( n \) nodes who can all communicate with each other. We start with a single node possessing a piece of information (“rumor”). The process works in discrete time steps (“rounds”). In each round, each informed node calls a node chosen uniformly at random (including itself) and gossips the rumor to it, making the node informed if it was not before.

This process is sometimes called synchronized rumor spreading in the push model in a complete graph. Note that this basic process also models the spread of a single piece of information in a system where several rumors are disseminated. Note also that the assumption that all nodes can communicate with each other makes sense even in networks where there is no physical connection between any two nodes. In such networks, gossip-based algorithms usually are implemented building upon a peer-sampling service [JVG+07], which enables the individual node to connect to a random other one.

1.3 Previous Results

For the basic rumor spreading process, several results exist. For simplicity, denote by \( S_n \) the number of rounds performed until (for the first time) all nodes are informed. This random variable is also called the broadcast time of the rumor spreading process.

Already Frieze and Grimmett [FG85] give the fairly precise result that \( S_n = (1 + o(1))(\log_2 n + \ln n) \) with probability \( 1 - o(1) \). They also prove that for all \( \varepsilon, \gamma > 0 \), with probability at least \( 1 - o(n^{-\gamma}) \), the broadcast time does not exceed \( (1 + \varepsilon)(\log_2 n + \gamma \ln n) \). The first bound was sharpened by Pittel [Pit87], who proved that for any \( h = \omega(1) \), we have \( \Pr(|S_n - \log_2 n - \ln n| \geq h(n)) \to 0 \).

1.4 Our Results

Despite the strong results of Frieze and Grimmett [FG85] and Pittel [Pit87], it is still surprising that a simple process like basic randomized rumor spreading is not even better understood. Recall, for example, that the coupon collector process is much better analyzed. We can precisely describe the time needed to collect all coupons as the sum \( C_n = X_1 + \ldots + X_n \) of \( n \) independent geometrically distributed random variables \( X_i \) with success rates \( p_i = i/n \). Consequently, the expected time the process takes is \( nH_n = n \ln(n) + \gamma n + 1/2 + O(1/n) \). For the basic rumor spreading process, we are not aware of any proof for an \( \log_2 n + \ln n + \Theta(1) \) bound for the expected runtime, let alone a precise description of the distribution. It might be possible to derive results in this direction from a careful analysis of the proof in [Pit87], but since the 8-page proof analyzing the process in 7 different phases is quite technical, we preferred to use an alternative route.

We prove that the expected time needed to inform all \( n \) nodes via the basic rumor spreading process is at most

\[
E[S_n] \leq \lceil \log_2 n \rceil + \ln n + 2.765 + o(1).
\]
In addition to the expectation, we show that the random variable describing rumor spreading time is dominated by

\[ S_n \preceq \lceil \log_2 n \rceil + 2.562 + o(1) + \frac{1 + O(n^{-\frac{1}{2} + \varepsilon})}{n} C_n + \text{Geom}(1 - O(n^{-1 + \varepsilon})) \],

where \( C_n \) is the time needed to collect \( n \) coupons, \( \text{Geom}(p) \) is an independent geometric random variable with success probability \( p \), and \( \varepsilon \) is an arbitrarily small constant.

These bounds are relatively sharp. For the expectation, a lower bound of \( \lceil \log_2 n \rceil + \ln n - 1.116 \) is not difficult to prove, see Section 4. Also, for the distribution, we show that \( S_n \) is subdominated by \( \lceil \log_2 n \rceil - 1 + (1/n) \sum_{i=1}^{n/2} X_i \), where as above the \( X_i \) are independent geometric random variables with success probability \( p_i = i/n \).

In addition to proving these relatively precise results, we also feel that our analysis method is slightly simpler than the previous works. We use the following two elementary, but powerful arguments.

(i) For each round of the process, we fix a target number of nodes that should become newly informed. We choose this number in such a way that the probability of failing the target is very small (at most \( n^{-1 + \varepsilon} \)). Consequently, by allowing an extra number of \( \text{Geom}(1 - n^{-1 + \varepsilon}) \) rounds, we can ensure that each such subphase reaches the target.

(ii) For the part of the process leading from \( cn \), \( c \) a constant, informed nodes to all nodes informed, we use an elegant reduction to the coupon collector process. We give (not too sharp) lower bounds for the number of informed nodes in these rounds, building on the elementary observation that the number of uninformed nodes typically shrinks by a constant factor. This factor is known to approach \( 1/e \), but we shall not exploit this and use a weaker factor. This weaker factor is enough to see that in the following \( \kappa \) rounds, a total of \( \kappa n - O(n) \) random contacts are made. From the coupon collector process, it is well-known how many random calls are needed to ensure that each of the missing \((1 - c)n\) receives a call.

We did not try to optimize the additive constants in our bounds, though we imagine that our proof method allows making them precise with moderate additional effort. We also note that our results confirms, and strengthens, previous observations that the first part of the process up to a constant fraction of informed nodes shows very little variation in the runtime. For example, we shall prove that with probability \( 1 - n^{-1 + \varepsilon} \), after \( \lceil \log_2 n \rceil - 1 \) rounds, at least \((5/16)n - o(n)\) nodes are informed. Note that within \( \log_2 n - 2 \) rounds, no rumor spreading process (where each informed node does one call per round) can inform more than \( n/4 \) nodes.

2 Preliminaries

Throughout this paper, we analyze the basic rumor spreading process introduced by Frieze and Grimmett [FG85] (see also [Pit87] for a beautiful description of the process). The process starts with one node of a complete network on the \( n \) nodes \([n] := \{1, \ldots, n\}\) knowing a rumor. In each round of the process, each node that knows the rumor chooses a node uniformly at random (including possibly itself) and gossips the rumor to that node, which becomes informed in case it was not before. We denote by \( I_t \) the set of informed nodes after the completion of round \( t \). Let \( I_0 \) consist of the single initially informed node. By \( U_t := [n] \setminus I_t \), we denote the set of uninformed nodes. Our main concern is the time needed to inform all nodes, that is, \( S_n := \min\{t \mid |I_t| = n\} \).
Before starting the analysis of this process, let us collect a few probabilistic tools needed in the following. Let a pair of random variables $X, Y$ be given. We say that $X$ stochastically dominates $Y$, and write $Y \preceq X$, if $\Pr[X \geq x] \geq \Pr[Y \geq x]$ for all $x$.

The following lemmas will be used to bound the deviation of random variables from their expectation, see, e.g., [DP09].

**Lemma 1** (Hoeffding’s bound for sums of independent random variables). Let $X := \sum_{i=1}^{n} X_i$, where the $X_i$ are independently distributed in $[0, 1]$. Then for all $t > 0$,

$$\Pr[X < E[X] + t] \leq e^{-2t^2/n}, \quad \Pr[X > E[X] + t] \leq e^{-2t^2/n}.$$ 

The previous lemma also holds in some dependent settings, in particular for negatively associated random variables. We say that random variables $X_1, \ldots, X_n$ are negatively associated, if for all disjoint subsets $I, J \subseteq [n]$ and all non-decreasing functions $f$ and $g$,

$$E[f(X_i, i \in I)g(X_j, j \in J)] \leq E[f(X_i, i \in I)]E[g(X_i, i \in I)].$$

**Lemma 2** (Hoeffding’s bound for sums of negatively associated random variables). Let $X := \sum_{i=1}^{n} X_i$, where the $X_i$ are negatively associated random variables taking values in $[0, 1]$. Then for all $t > 0$,

$$\Pr[X < E[X] + t] \leq e^{-2t^2/n}, \quad \Pr[X > E[X] + t] \leq e^{-2t^2/n}.$$ 

We say that a random variable $G$ is geometrically distributed with success probability $p$, and write $G \sim \text{Geom}(p)$, if $\Pr[G = i] = (1 - p)^i p$ for all $i \in \mathbb{N}$. Note that we allow $G$ to take the value zero. We will typically bound the expected number of rounds to inform a certain number of nodes by a deterministic number and a sum of geometrically distributed variables. This value is typically dominated by the single geometric random variable with the smallest success probability, which we will use to simplify the results. The following lemma helps to make this observation formal.

**Lemma 3.** Let $G_1, \ldots, G_n$ be independent random variables with $G_i \sim \text{Geom}(1 - q_i)$. Then $\sum_{i=1}^{n} G_i$ is stochastically dominated by a random variable $G$ with $G \sim \text{Geom}(1 - \sum_{i=1}^{n} q_i)$.

**Proof.** By induction (details omitted). \qed

### 3 Upper Bounds

To prove the upper bounds, we use three different ways to prove lower bounds on the numbers of informed vertices. When there are few informed vertices, a birthday paradox type computation shows that very likely, all calls reach different uninformed vertices (Phase 1). When there are more, but at most a small constant fraction of informed nodes, a similar argument together with a Chernoff bound argument shows that the number of informed nodes almost doubles (Phase 2). When there are even more informed nodes, then, like previous works, switching the focus to the uninformed nodes and again using Chernoff bounds shows that the number of uninformed nodes shrinks by a constant factor (Phase 3). This shows that in the following $p_3$ rounds, $p_3n - O(n)$ random calls are made in total. Together with a reduction to the coupon collector process, we derive a bound on the rumor spreading time.
3.1 Phase 1

When only few nodes are informed, the random calls performed by these nodes have a very high chance of targeting an uninformed node, and consequently, there is a good chance that the number of informed nodes doubles. The following lemma makes this observation precise.

**Lemma 4.** Let $n_1 < \sqrt{n}$ be a power of two and $t_1 := \min\{t \geq 0 \mid |I_t| \geq n_1\}$. Then $t_1$ is stochastically dominated by $\log_2(n_1) + \text{Geom}(1 - n_1^2/n)$.

**Proof.** Let $p_1 := \log_2 n_1$, $2 \leq j \leq p_1$ and $t \geq 0$. Assume that in the $(t + 1)$st rounds, the $|I_t|$ informed nodes perform their actions in some given order. Then, when the $k$th node chooses its random communication partner, at most $|I_t| + k - 1$ nodes are informed. Consequently, with probability at least $1 - \left(\frac{|I_t| + k - 1}{n}\right)^2$, it calls an uninformed node and informs it. Hence

\[
\Pr[|I_{t+1}| \geq 2^j \mid |I_t| \geq 2^{j-1}] \geq \prod_{k=0}^{2^{j-1}-1} \left(1 - \frac{2^j - 1 + k}{n}\right) \geq 1 - \frac{2^{2(j-1)} + \sum_{k=0}^{2^{j-1}-1} k}{n} \geq 1 - \frac{2^{2(j-1)} + 2^{2(j-1)-1}}{n},
\]

where the second inequality follows from Bernoulli’s inequality.

We divide the process until at least $n_1$ nodes are informed into $p_1$ subphases, where each subphase $j$ ends as soon as at least $2^j$ nodes are informed. These subphases almost surely consist of a single round each. Formally, define the random stopping times

\[t^{(j)} := \min\{t \geq 0 \mid |I_t| \geq 2^j\}, \text{ for } 1 \leq j \leq p_1.\]

Clearly, at time $t^{(p_1)} = t^{(1)} + \sum_{j=2}^{p_1} t^{(j)} - t^{(j-1)}$ at least $2^{p_1} \geq n_1$ nodes are informed. Let $2 \leq j \leq p_1$. The above calculation shows that $t^{(j)} - t^{(j-1)}$ is stochastically dominated by $1 + \text{Geom}(1 - q_j)$ with $q_j := (2^{2(j-1)}) + 2^{2(j-1)-1})/n$, since in each round $t$ with $|I_t| \geq 2^{j-1}$, we have that $|I_{t+\ell}| < 2^j$ with probability at most $q_\ell$ for each $\ell \geq 1$. Similarly, $t^{(1)}$ is stochastically dominated by $1 + \text{Geom}(1 - q_1)$ with $q_1 := 1/n$.

We introduce independent random variables $G_j \sim \text{Geom}(1 - q_j)$ and obtain

\[t^{(p_1)} \leq p_1 + \sum_{i=1}^{p_1} G_i.\]

To complete the proof, note that

\[\sum_{j=1}^{p_1} q_j \leq \frac{1}{n} + \sum_{j=2}^{p_1} \frac{2^{2(j-1)} + 2^{2(j-1)-1}}{n} \leq \frac{2^{p_1}-1}{n},\]

and apply Lemma 3. This proves the claim. \qed
3.2 Phase 2

When more than just very few nodes are informed, the number of informed nodes stops doubling with high probability in one round. However, the number of informed nodes increases, in expectation, almost by a factor of two per round, and is additionally closely concentrated around its expectation. This is what we make precise in this subsection.

Let \( 1 \leq k \leq \frac{n}{2} \) and assume that \( |I_t| = k \). Enumerate \( I_t \) in an arbitrary manner \( u_1, \ldots, u_k \). Denote by \( c(u_j) \) the random node called by \( u_j \) in the \((t+1)\)st round. Let \( X_j \) be the indicator random variable for the event \( c(u_j) \notin I_t \cup \{c(u_1), \ldots, c(u_{j-1})\} \). Then \( |I_{t+1}| = k + \sum_{j=1}^{k} X_j \). While the \( X_j \) are not independent, they satisfy the property that regardless of the outcome of \( X_1, \ldots, X_{j-1} \), we have \( \Pr[|X_j| = 1 \geq 1 - (k + j - 1)/n] \). This not only allows us to bound the expectation of \( |I_{t+1}| \) by

\[
E[|I_{t+1}|] \geq 2k - \frac{k^2}{n} + \frac{k-1}{n} \geq 2k - \frac{3k^2}{2n} = 2k \left(1 - \frac{3k}{4n}\right),
\]

but also allows to use Hoeffding bounds. The above property implies that \( X_1 + \ldots + X_k \) stochastically dominates \( Y_1 + \ldots + Y_k \), where the \( Y_j \) are independent binary random variables with \( \Pr[Y_j = 1] = 1 - (k + j - 1)/n \). This fact seems to be well known, but the only published proof we are aware of is Lemma 1.18 in [Doe11].

**Lemma 5.** Let \( 1 \leq k \leq \frac{n}{2} \) and \( \delta > 0 \), then

\[
\Pr \left[ |I_{t+1}| \leq 2k \left(1 - \frac{3k}{4n} - \frac{1}{k^{2-\delta}}\right) \right| |I_t| \geq k \leq e^{-4k^{2\delta}}.
\]

**Proof.** On the event \( |I_{t+1}| \leq 2k \left(1 - \frac{3k}{4n} - \frac{1}{k^{2-\delta}}\right) \), we have

\[
\sum_{i=1}^{k} X_i \leq k \left(1 - \frac{3k}{2n} - \frac{2}{k^{2-\delta}}\right) \leq E \left[ \sum_{i=1}^{k} X_i \right] - 2k^{1+\delta}.
\]

Hoeffding’s bound (Lemma 1) completes the proof by bounding

\[
\Pr \left[ \sum_{i=1}^{k} X_i \leq E \left[ \sum_{i=1}^{k} X_i \right] - 2k^{1+\delta} \right] \leq e^{-4k^{2\delta}}.
\]

\[\square\]

**Lemma 6.** Let \( k \in \mathbb{N}, k \geq 1 \), \( s > 0 \) and \( 0 < \delta < s/2 \). Define \( n_2 := n \left(\frac{1}{2^s} - \frac{3}{4} \cdot \frac{1}{2^s} - \epsilon\right) \), and correspondingly, \( t_2 := \min \{t \geq 0 \mid |I_t| \geq n_2\} \). Then \( t_2 \) is stochastically dominated by

\[
[\log_2(n)] - k + \text{Geom}(1 - (n^{-1+\delta} + \log_2(n)e^{-(n^{s/2})^{2\delta}})).
\]

**Proof.** We define \( n_1 \) as largest power of two that is at most \( n^s \). We apply Lemma 4 to analyze \( t_1 := \min \{t \geq 0 \mid |I_t| \geq n_1\} \), and introduce, analogously to the proof of Lemma 4, subphases \( t^{(1)}, t^{(2)}, \ldots \) with

\[
t^{(j)} := \min \{t \geq 0 \mid |I_t| \geq a_j\},
\]

6
where \( a_0 := n_1 \) and \( a_i := 2a_{i-1}(1 - \frac{3a_{i-1}}{4n} - a_{i-1}^{-(1/2)+\delta}) \). Trivially, \( a_i \leq 2^{i}a_0 \) and \( a_i \geq a_0 \). Hence,

\[
a_i \geq 2a_{i-1} \left( 1 - \frac{3}{4} 2^{i-1}a_0 - \frac{1}{a_0^{1/2-\delta}} \right) \geq 2^{i}a_0 \prod_{j=1}^{i} \left( 1 - \frac{3}{4} 2^{j-i}a_0 - \frac{1}{a_0^{1/2-\delta}} \right)
\]

\[
\geq 2^{i}a_0 \left( 1 - \frac{3}{4} 2^{i}a_0 - \frac{i}{a_0^{1/2-\delta}} \right).
\]

Note that after \( p_2 := \lfloor \log_2(n) \rfloor - k - \log_2(n_1) \) subphases, we have

\[
a_{p_2} \geq \frac{n}{2^k} \left( 1 - \frac{3}{4} 2^{k} - \frac{\log_2(n)}{n^{1/2-\delta}} \right) \geq n_2,
\]

for sufficiently large \( n \).

Consequently, \( t_2 \) is dominated by

\[
t_1 + \lfloor \log_2(n) \rfloor - k - \log_2(n_1) + \sum_{i=1}^{p_2} G_i, \text{ where } G_i \text{ is geometrically distributed with success probability } 1 - q_i \text{ with } q_i \leq e^{-4n_1^2i^2} \text{ using Lemma 5}. \]

Since Lemma 4 yields

\[
t_1 \leq \log_2(n_1) + \text{Geom} \left( 1 - \frac{n^2}{n} \right), \text{ and } \sum_{i=1}^{p_2} q_i \leq \log_2(n)e^{-4n_1^2} \leq \log_2(n)e^{-4(n/2)^2}, \text{ we obtain, by Lemma 3,}
\]

\[
t_2 \leq t_1 + \lfloor \log_2(n) \rfloor - k - \log_2(n_1) + \text{Geom}(1 - \log_2(n)e^{-4(n/2)^2}) + \sum_{i=1}^{p_2} q_i \leq \lfloor \log_2(n) \rfloor - k + \text{Geom}(1 - (n^{-1/2} + \log_2(n)e^{-4(n/2)^2})).
\]

\[
\square
\]

### 3.3 Phase 3

Once a linear number of nodes is informed, the probability for a specific uninformed nodes to stay uninformed is less than a constant, hence we switch our focus from the set of informed nodes to the set of uninformed nodes. Let \( |U_t| \leq cn \) for some constant \( 0 < c < 1 \). Then

\[
\mathbb{E}[|U_{t+1}|] \leq |U_t| \left( 1 - \frac{1}{n} \right)^{|I_t|} \leq cn \left( 1 - \frac{1}{n} \right)^{(1-c)n} \leq cne^{-(1-c)}.
\]

In fact, the number of uninformed nodes is concentrated in an \( O(n^{(1/2)+\delta}) \)-interval around its expectation.

**Lemma 7.** For any \( \delta > 0 \), we have

\[
\Pr \left[ |U_{t+1}| > cne^{-(1-c)} + n^\frac{1}{2} + \delta \right] \leq cn \leq cne^{-(1-c)}.
\]

**Proof.** W.l.o.g. assume that \( U_t = \{1, \ldots, u\} \) with \( u \leq cn \). We introduce \( X := \sum_{i=1}^{cn} X_i \), where we set the indicator variables \( X_i = 1 \) if and only if node \( i \) is not informed by any node in \( I_t \). The indicator variables are negatively associated (see, e.g., chapter 3 in [DP09]) and \( \mathbb{E}[X] \leq cne^{-(1-c)} \). Note that node \( i \) is not necessarily uninformed, however, \( |U_{t+1}| \leq X \) holds in any case. Consequently,

\[
\Pr \left[ |U_{t+1}| > cne^{-(1-c)} + n^\frac{1}{2} + \delta \right] \leq \Pr \left[ X > \mathbb{E}[X] + n^\frac{1}{2} + \delta \right] \leq e^{-2n^{\frac{1}{2}+\delta}}.
\]

using Lemma 2. \( \square \)
Lemma 8. Let $0 < c < 1$ and $t^{(0)} := \min\{t \geq 0 \mid |U_t| \leq cn\}$. For any $\delta > 0$, the number of rounds until all nodes are informed is stochastically dominated by

$$t^{(0)} + (1 + g(n)) \frac{C_n(cn)}{n} + \frac{c}{1 - e^{-(1-c)}} + h(n) + \text{Geom}(1 - e^{-n\delta}),$$

where with $q := e^{-(1-c)}$ and $Z := \left(c + \frac{1}{1-q}\right)n^{-\frac{1}{2}+\delta}$, we define

$$g(n) := \frac{Z}{1 - Z} = O\left(n^{-\frac{1}{2}+\delta}\right)$$

$$h(n) := \frac{(1/2 - \delta)n^{-\frac{1}{2}+\delta} \ln n}{(1-c)(1-q)} + g(n) \left(\frac{c}{1-q} + \frac{(1/2 + \delta)n^{-\frac{1}{2}+\delta} \ln n}{(1-c)(1-q)}\right) = O\left(n^{-\frac{1}{2}+\delta} \ln n\right).$$

For some $p_3$ to be chosen later, we introduce $p_3$ subphases almost surely consisting of single rounds each, as in the previous phases. To this end, we set $\varepsilon := n^{-1/2+\delta}$, $z_0 := c$ and $z_i := z_{i-1}e^{-(1-z_{i-1})} + \varepsilon$. Let $t^{(i)} := \min\{t > t^{(i-1)} \mid |U_t| \leq z_i n\}$ be the time that concludes the $i$-th subphase of Phase 3. Using the previous lemma, we immediately see that

$$\Pr[t^{(i)} - t^{(i-1)} > 1] \leq \Pr[|U_{t+1}| > z_i n \mid |U_t| \leq z_{i-1} n] \leq e^{-2n^{2\delta}}.$$

Consequently, using Lemma 3, $\sum_{i=1}^{p_3}(t^{(i)} - t^{(i-1)}) \leq p_3 + \text{Geom}(1 - p_3 e^{-n^{2\delta}})$.

Observe that $z_i \leq c$ for all $i$ and thus, with $q := e^{-(1-c)}$,

$$z_i \leq z_{i-1}q + \varepsilon \leq \cdots \leq q^i c + \sum_{j=0}^{i-1} q^j \varepsilon.$$

Clearly, the total number of messages sent in the subphases $1, \ldots, p_3$ is lower bounded by

$$\sum_{i=0}^{p_3-1} n(1 - z_i) = n \left(p_3 - \sum_{i=0}^{p_3-1} z_i\right) \geq n \left(p_3 - c \sum_{j=0}^{p_3-1} q^j \varepsilon p_3 \sum_{j=0}^{p_3-1} q^j\right) \geq n \left(p_3 - \frac{c}{1-q} - \frac{p_3 \varepsilon}{1-q}\right).$$

Let $C_n(x)$ be a random variable denoting the number of coupons to be drawn in the coupon collector process with $n$ coupons until a set of $x$ distinguished coupons has been collected. Equivalently, this is the time needed to collect all coupons given that we already start with $n - x$ distinct coupons. We couple the rumor spreading process and the coupon collector process by identifying informed nodes and coupons, and mimicking the choice of message receivers as coupons in the coupon collector process. We observe that if the number of messages sent after round $t^{(0)}$ is larger than $C_n(cn)$, the rumor spreading process is completed.

Note that when collecting all coupons, i.e., nodes receiving the rumor, in the subphases $1, 2, \ldots, p_3$, by definition it is assured that at least $cn - z_{p_3}n$ of the distinguished coupons appear among these coupons. Hence these are not independent draws from the set of all coupons. However, the following basic fact (proof omitted) shows that this precondition only decreases the number of remaining draws until all coupons are collected.
Lemma 9. Let $C_1, C_2, \ldots$ be a sequence of coupons drawn independently at random from $[n]$, let $X \subseteq [n]$ be a set of $x$ distinguished coupons and $M$ be arbitrary. We define $S \subseteq X$ as the set of coupons of $X$ collected among $C_1, \ldots, C_M$. Then, for any $0 \leq m \leq M$ and $r$,

$$
\Pr[C_n(x) \leq r \mid |S| \geq m] \geq \Pr[C_n(x) \leq r],
$$
i.e., $C_n(x)$ conditioned on $|S| \geq m$ is stochastically dominated by $C_n(x)$.

Consider the situation in round $t^{(p_3)}$ and afterwards. Setting $p_3 := \log q \epsilon$, the number of uninformed nodes is bounded by

$$
z_{p_3} \leq cq^{\log q \epsilon} + \epsilon \sum_{j=0}^{p_3-1} q^j \leq \left( c + \frac{1}{1-q} \right) \epsilon.
$$

Consequently, in each additional round $n - O(\epsilon)$ messages are sent. Let $C = C_n(cn)$ conditioned on the coupons collected in rounds $t^{(0)}, \ldots, t^{(p_3)}$ fulfilling the assertions of the subphases and $j$ be such that

$$
j(n - z_{p_3}n) + n \left( p_3 - \frac{c}{1-q} - \frac{p_3 \epsilon}{1-q} \right) = C,
$$
i.e, the number of messages sent in rounds $t^{(0)} + 1, \ldots, t^{(p_3)} + [j]$ is at least the required amount of $C$ and the process is completed. We compute

$$
j = \frac{n}{n - z_{3n}} \left( \frac{C}{n} + \frac{c}{1-q} + \frac{\epsilon p_3}{1-q} - p_3 \right)
\leq \frac{C}{n - z_{3n}} + \frac{c}{1-q} - p_3 + \frac{\epsilon p_3}{1-q} + \frac{z_{p_3} + \epsilon p_3}{1-q}
= \frac{C}{n - z_{3n}} + \frac{c}{1-q} - p_3 + h_c(\epsilon),
$$

where $h_c(\epsilon) := \frac{\epsilon \log q \epsilon}{1-q} + \frac{z_{p_3} + \epsilon \log q \epsilon}{1-q}$.

By repeated application of Lemma 9, we can replace $C$ by the independent random variable $C_n(cn)$ and conclude that the total number of rounds is stochastically dominated by

$$
t^{(0)} + \sum_{i=1}^{p_3} (t^{(i)} - t^{(i-1)}) + [j] \leq t^{(0)} + \text{Geom}(1 - p_3 e^{-n^{2s}}) + \left[ \frac{C_n(cn)}{n - nz_3} + \frac{c}{1-q} + h_c(\epsilon) \right].
$$

3.4 Connecting the Phases

Theorem 10. Let $s > 0$, $0 < \delta < s/2$. The number of rounds until the rumor spreading process on the complete graph with $n$ nodes informs all nodes is stochastically dominated by

$$
[\log_2 n] + \left[ (1 + g(n)) \frac{C_n((\frac{11}{10} + \epsilon) n)}{n} + 1.562 + 10.7 \epsilon + O(\epsilon^2) + h(n) \right]
+ \text{Geom} \left( 1 - \left( \frac{1}{n^{1-s}} + \log_2(n) e^{-(n^{s/2})^{2\delta}} + e^{\epsilon \delta} \right) \right),
$$

where $\epsilon := \frac{\log_2(n)}{n^2 s - \delta}$, $g(n) = O(n^{-(1/2) + \delta})$ and $h(n) = O(n^{-(1/2) + \delta} \ln n)$.
Proof. By Lemma 6, we have that for \( \bar{c} := \frac{5}{16} - \varepsilon \), the time until at least \( \bar{c}n \) nodes are informed is stochastically dominated by \( \lceil \log_2 n \rceil - 1 + \text{Geom}(1 - (n^{-1+s} + \log_2(n)e^{-(n^s/2)\delta^4})) \). Set \( c := 1 - \bar{c} = \frac{11}{16} + \varepsilon \) and \( F(x) := \frac{x}{1-e^{-1-x}} \). Then \( F(c) \leq F(11/16) + F'(11/16)\varepsilon + o(\varepsilon^2) \leq 2.562 + 10.7\varepsilon + o(\varepsilon^2) \).

Lemma 8 yields that the total number of rounds is stochastically dominated by

\[
\lceil \log_2 n \rceil + \left[ 1 + g(n) \frac{C_{e(cn)}}{n} + F(c) + h(n) \right]
+ \text{Geom} \left( 1 - \left( n^{-1+s} + \log_2(n)e^{-(n^s/2)\delta^4} \right) \right) + \text{Geom}(1 - e^{n\delta}).
\]

Using that \( E[C_{n}(cn)/n] \leq H_{cn} \leq \ln(n) - \ln(c) + \gamma + O(1/n) \), we immediately derive the following statement, since the expected number of total rounds is bounded by

\[
\lceil \log_2 n \rceil + \ln(n) - \ln(11/16) + \gamma + 1.562 + 1 + o(1).
\]

Corollary 11. The expected number of rounds until the rumor spreading process on the complete graph with \( n \) nodes informs all nodes is at least

\[
\lceil \log_2 n \rceil + \ln(n) - 0.116.
\]

Proof. Using \( H_x \geq \ln(x) + \gamma - O(1/n) \), we compute

\[
E \left[ \frac{C_{n}(n/2)}{n} \right] \geq \ln(n) + \gamma - \ln(2) \geq \ln(n) - 0.116.
\]
References


