On Collaborative Content Distribution using Multi-Message Gossip

Coby Fernandess[*] Dahlia Malkhi[†]

Abstract

We study epidemic schemes in the context of collaborative data delivery. In this context, multiple chunks of data reside at different nodes, and the challenge is to simultaneously deliver all chunks to all nodes.

Here we explore the inter-operation between the gossip of multiple, simultaneous message-chunks. In this setting, interacting nodes must select which chunk, among many, to exchange in every communication round.

We provide an efficient solution that possesses the inherent robustness and scalability of gossip. Our approach maintains the simplicity of gossip, and has low message, connections and computation overhead. Because our approach differs from solutions proposed by network coding, we are able to provide insight into the tradeoffs and analysis of the problem of collaborative content distribution. We formally analyze the performance of the algorithm, demonstrating its efficiency with high probability.

1 Introduction

Collaborative content delivery is at the focus of tremendous recent attention, driven by the growing need for applications such as file sharing, web cast, software distribution, etc. A collaborative multicast is initiated by breaking the content into *chunks*, each one sent to a different node (or set of nodes). Subsequently, the nodes exchange the chunks they hold among themselves until each node collects copies of all the chunks. The advantage of the collaborative approach is obvious. Because a source may become choked with a *flash crowd* of demanding clients, clients at the endpoints cooperate in delivering the content. This alleviates the bottleneck at the content distributor, and provides total bandwidth that scales with the number of participants.

The most widely deployed content delivery systems on the Internet today, BitTorrent and Emule, operate this way. In these systems, as soon as a client obtains a chunk, it becomes a download source for forwarding that chunk. In many ways, this distribution process resembles a randomized gossip process with multiple origin points: Clients select random partners among the set of current downloaders and exchange chunks with them.

However, experience with BitTorrent and similar systems indicates that the main problem with this approach is that towards the end of a download, many peers may be missing the same 'rare' chunks, and the download slows down. In a lot of ways, this core difficulty resembles that of the famous coupon collector problem, and in both, much of the complexity results from the finishing steps.

[*] School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel. fery@cs.huji.ac.il

[†] School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel and Microsoft Research, Silicon Valley Campus. dalia@microsoft.com

This paper takes a formal view of collaborative exchange of multiple data items in a network of nodes using gossip techniques. Consider the problem in a somewhat more formal manner now (precise definitions are given in Section 2). In (semi-) synchronous rounds, each node selects a partner uniformly at random, and exchanges chunks with it. The selection of chunks is the crucial point of investigation in this paper. In Emule [KB05], for example, a node transfers to its partner a chunk selected uniformly at random among those missed by the partner. It is easy to show that under this strategy, with non-negligible probability some chunks will initially spread more quickly than others. This process intensifies itself, since the more sources there are for a chunk, the faster it spreads. Due to the exponential nature of spreading, the process may choke some chunks, leading to very slow, even linear, dissemination time of these chunks.

Recently, *network coding* was suggested as a means to alleviate this problem. In this approach, rather than distributing k different chunks of the data over different paths, a randomized linear combination of the initial chunks is sent to each destination. Once a node obtains these re-encoded chunks, it can generate new combinations from the ones it has, and send those out to other nodes. The main benefit is that nodes can make use of any new chunk if it is linearly independent from previous ones. It can be shown that independence of randomized linear combination is achieved in most cases. This means that no one node can become a bottleneck, since no specific combination is more important than any other. Once a node collects sufficient independent combination-chunks, it may use them to reconstruct the whole content.

Several deployments of this idea have already emerged [JLC05, GR05]. A formal analysis of randomized content exchange using network coding in the case where the initial set size, k, is $\Theta(n)$ was given in [DM04]. It shows asymptotically optimal running time of $O(n)$ rounds *with high probability* (w.h.p.) [1]. One of the drawbacks of this approach is its relatively high message payload overhead, namely, $O(n \ln n)$ bits per message. A theory is put forward in [DM04] that in some sense, this cost is mandatory for efficiency. More concretely, the conjecture is that any store-and-forward protocol (without manipulation of messages) cannot converge in less than $O(n \ln n)$ rounds.

Our result is a gossip protocol that provides insight into the tradeoffs and analysis of spreading any initial set size of k chunks among all nodes in $O(k + \ln n)$ rounds. This round complexity is asymptotically optimal: The dissemination of a single chunk requires at least $\ln n$ rounds. On the other hand, the dissemination of k chunks to n nodes, where in each round no more than n chunks are sent over the network connections, requires at least $\frac{kn}{n} = k$ rounds. Hence the running time is lower bounded by $max\{k, \ln n\} = \Omega(k+\ln n)$. Furthermore, when $k \geq \ln n$, our protocol makes use of a total of $O(kn)$ network connections, which is asymptotically optimal. Our protocol serves to disprove the conjecture by Medard et al. [DM04]: It is a store-and-forward protocol that achieves an asymptotically optimal running time of $O(k + \ln n)$ rounds w.h.p. for any initial set size k. In particular, it requires $O(n)$ rounds when $k = O(n)$.

Our work bears significance on the fundamental theory of gossip networks. Gossip is a powerful paradigm in distributed computing. Gossip protocols spread messages (chunks) obliviously, without centralized control or management, with remarkable speed and with inherent fault tolerance. Epidemic-style gossip techniques for information dissemination are central in numerous distributed systems, e.g., Usenet news [LOM94], the Grapevine distributed system [BLNS82], Ad Hoc routing [HHL02], distributed failure detectors [vRMH98], the Astrolabe network management system [vR00], lightweight broadcast [EGH$^+$03], membership maintenance [GKM03], GosSkip [GHK04], the CYCLON system [VGvS05], and others.

[1] with high probability (w.h.p.), meaning that the claim holds with probability of at least $1 - O(\frac{1}{n})$.

The gossip process of a single message has been investigated for more than a decade [SHL88, DGH$^+$87, KKD04, KSSV00]. However, the investigation of multiple, simultaneous gossip messages dissemination is still in its early stages. Several previous works addressed multi-source gossip experimentally, *e.g.*, [BHOO$^+$99, EGH$^+$03, SS00]. From a formal point of view, if a node could forward all the gossip messages it has obtained in one step, then multi-messages gossip can be seen as an immediate extension of single-message gossip. However, sending large content in this manner is wasteful, and defies the whole purpose of breaking the file into chunks in order to avoid repeated store-and-forward. Our results are the first to shed light on the simultaneous gossip of multiple messages under a bandwidth constraint that allows one message transfer per round.

To summarize, we provide the following contribution:

1. We present the first formal study of multi-message store-and-forward gossip protocols under bandwidth constraints.

2. We provide a clean multi-message gossip protocol that exhibits asymptotically optimal behavior under bandwidth constraints. Specifically, the protocol spreads k initial messages among n participants in $O(k + \ln n)$ bandwidth-limited rounds, using $O(n(k + \ln n))$ two-way connections.

3. Our study serves to refute a recent conjecture by Medard et al. [DM04] concerning the time lower bound for store-and-forward spreading of data.

4. An immediate consequence of our gossip protocol is an efficient, gossip-based collaborative content delivery mechanism. Our protocol overcomes the "rarest chunks" problem and avoids unnecessary delay using simple means and does not employ coding.

Technical Approach. Our approach stems from the following key observation: In order to provide efficient multi-message dissemination, messages must spread equally wide (roughly). To this end, our protocol employs a *coloring* mechanism whereby each node has a unique color that indicates the message for which it has primary forwarding responsibility. We devise a simple distributed *aging* mechanism that limits the scope of primary messages dissemination to $n/2k$ nodes. As a result, the total number of colored nodes does not exceed $n/2$, and no color becomes choked. In the final k rounds of the protocol, we use a message exchange policy that determines which message to exchange with a gossip partner based on the nodes' respective colors, their ages, and randomization.

Organization. The paper is organized as follows: System model, definitions and preliminaries are given in Section 2. Section 3 describes our main gossip based protocol, which achieves an asymptotically optimal running time w.h.p., and its analysis. Section 5 describes a variant of our protocol with reduced payload. Finally, in Section 6 we review related work and our conclusion is given in Section 7.

2 Definitions and Preliminaries

The system consists of a set V of n processor nodes interconnected by a complete graph (clique). Each node has a unique identifier, $i \in [n]$. The general structure of all the gossip protocols discussed in this paper is that of the *anti-entropy* protocol of Demers et al. [DGH$^+$87]. These protocols operate in synchronous rounds, denoted $r = 1, 2, \ldots$. In each round r, every node v chooses a communication partner $u \in V$ at random, and they exchange messages in order to resolve their differences. With reference to the flow of information, [DGH$^+$87] has distinguished between push and pull transmission models. Assume node v calls node u.

- The message is *pushed* if v transfers u a message.
- The message is *pulled* if u transfers v a message.

We enforce a strict *connectivity bound* as follows. In one round, a processor may initiate exactly one outgoing connection, and receive at most one. Every connection may carry at most one message, plus any payload prescribed by the protocol.

Problem statement. In this paper, we investigate multiple-message gossip protocols. We note that our original motivation is to spread *chunks* belonging to one data object, but in our formulation, we simply call them messages. Our problem statement is as follows. Let $I \subseteq V$ denote the initial set of processors in V that hold the messages, $|I| = k \leq n$, which they wish to disseminate to all other nodes.

Naturally, the goal is to efficiently disseminate all the messages among all network nodes. More concretely, efficiency is manifested in the following criteria:

Time Total number of rounds for delivery, as measured from when the initial messages are generated and until all the messages are delivered at all the nodes w.h.p.

Communication complexity The total amount of data transferred over the network connections.

A node that wishes to disseminate a new message must do so in a predefined time slot. In other words, we do not consider continuous injection of new messages into the system. Additionally, we assume that n and k are known to the participants, although in Section 4 we show that given n, k can be automatically obtained in $O(\ln n)$ initial rounds using the techniques of [KDG03].

Slightly abusing notation, we identify each message by the processor $i \in I$ it was originated from. We denote by $M_v(r)$ the set of messages stored by v at the beginning of round r. When $M_v(r)$ is sent in a message, this is done by sending an n-bit vector in $\{0, 1\}^n$, indicating the presence of messages in the set according to their indices. A node v may obtain during the course of the protocol a color, denoted $c_v \in I$. The color of a node is the index of the first message v received. Each colored node stores its color message along with a time-varying integer value called *age*, denoted a_v, which bears resemblance to the aging technique used in [BHOO$^+$99]. For an origin processor $i \in I$, the initial color equals to $c_i = i$ along with initial age $a_i = 0$. Color messages may be gossiped; we will demonstrate in the protocol below the rules for maintaining and gossiping the message age. For a colored node $v \in V$, we use the notation $\langle c_v, a_v \rangle$ to denote the message c_v along with its age a_v. A node with no color message is *uncolored*. Last, we denote by $C(i, r)$ the set of nodes with color $i \in I$ at the beginning of round r, and by $c(i, r) = |C(i, r)|$ its size.

3 A Time Optimal Multi-Message Gossip Protocol

In this section, we present the main result of this paper, a gossip protocol whose time complexity is asymptotically optimal. The protocol employs a pull-based strategy, with several key components. First, we introduce a *coloring* mechanism, whereby each node has a unique color that indicates the message for which it has primary forwarding responsibility. Second, we devise an *aging* mechanism that limits the scope of primary message dissemination. Finally, we devise a message exchange policy that determines which message to exchange with a gossip partner based on the nodes' respective colors, their ages, the messages they hold and randomization.

3.1 The Protocol

In each round $r > 0$, every node u pulls a gossip partner v uniformly at random. Node u sends v a pull request, which contains the bit-vector $M_u(r)$, and an indication whether it has a color or not. Node v acts as follows:

1. If v has color c_v, u does not have a color, and c_v's age, a_v, satisfies $a_v < \log \frac{n}{2k}$, then v increments c_v's age to $a_v + 1$, and sends the color message along with its updated age, $\langle c_v, a_v \rangle$, to u.

2. Otherwise, if node v's color message $c_v \in M_v(r) \setminus M_u(r)$, then v sends c_v to u.

3. Otherwise, v selects a message from $M_v(r) \setminus M_u(r)$ at random and sends it to u.

When u gets the response from v, it stores it locally. If this is a color message with an age, then u becomes colored with color c_v and stores its age along with it.

3.2 Protocol Analysis

Our analysis shows a sharp termination in $O(\ln n + k)$ rounds. That is, throughout the analysis we concentrate on demonstrating that our results hold *with high probability* (w.h.p.), namely, with probably at least $1 - \frac{1}{n}$. Before going into the details of the proof and the related analysis, we provide the key intuition behind our protocol analysis. We have divided our analysis into three distinct chronological phases, each is represented by one case in the protocol's exchange policy.

1. By the end of the first phase the number of colored nodes of every colored message is exactly $\frac{n}{2k} - 1$ w.h.p.

2. By the end of the second phase the probability that each node has obtained any specific color message is at least $\frac{1}{2}$.

3. Last, all the messages are delivered to all the nodes w.h.p.

We begin with a technical lemma that contains the effect of bounded connectivity. We say that a round r is a *collision-free* round for a node v if during round r, v pulls a node that is not pulled by any other node in round r.

Lemma 3.1 *For every $t \geq 48 \ln n$ consecutive rounds there are at least $t/6$ rounds that are free of collisions for all nodes $v \in V$ w.h.p.*

Proof. Fix any $v \in V$. Let $z_v(r)$ be an independent Bernoulli trial that equals 1 if round r is a collision-free round for node v, 0 otherwise. The probability $z_v(r)$ satisfies:

$$p(z_v(r) = 1) = \left(1 - \frac{1}{n}\right)^{(n-1)} \geq \frac{1}{e} \ .$$

Let $Z_v = \sum_{r=r_0 \ldots r_0+t} z_v(r)$ be the number of collision-free rounds for v. Then $\mu = E[Z_v] \geq \frac{t}{e}$. Applying a Chernoff bound, we obtain with an appropriate choice of constants:

$$\begin{aligned} Pr[Z_v \leq t/6] &< Pr[Z_v \leq (1-\delta)\mu] \\ &\leq \exp\left(\frac{-\mu\delta^2}{2}\right) \\ &< n^{-2} \end{aligned}$$

Where $t \geq 48 \ln n$, $\mu \geq t/e > t/3$, and $\delta = 1/2$.

Applying union bound, we conclude that there are at least $t/6$ rounds that are collision-free for all nodes $v \in V$ w.h.p. □

Lemma 3.2 *(Phase 1) For any given run R of $O(\ln n)$ rounds of our gossip algorithm, the number of colored nodes of every colored message is exactly $\frac{n}{2k} - 1$ w.h.p.*

Proof. (Sketch) Since every color message must branch once in order for its age to increase, we obtain from this an upper bound of $\frac{n}{2k} - 1$ on the total number of nodes with the same color in the system. In order to complete our proof, we show that after $O(\ln n)$ rounds the size of each color set is $\frac{n}{2k} - 1$ nodes w.h.p. Let $p(r)$ denote the probability that a colored node is pulled by an uncolored one during round r. Since half the nodes are uncolored at any time, $p(r)$ is always lower bounded by $\frac{1}{3}$:

$$p(r) \geq 1 - (1 - \frac{1}{n})^{\frac{n}{2}} \geq 1 - e^{-\frac{1}{2}} > \frac{1}{3}.$$

As before, since every color message must branch once in order for its age to be increased, obtain that after $r = 3 \log \frac{n}{2k}$ rounds the expectation of a color-age a_v stored at any colored node v is $\log \frac{n}{2k}$. Obtaining this as a high probability result is considerably more involved, and is deferred to the appendix. Consequently, each color set $C(i, r)$ corresponds to a disjoint tree rooted at $i \in I$ of $\frac{n}{2k}$ nodes as required. □

Lemma 3.3 *(Phase 2) For every node $v \in V$ and for all colors $i \in I$, round $r = 18 \ln n + 48 \ln n + 12k$ satisfies $Pr[i \in M_v(r)] > \frac{1}{2}$.*

Proof. Following our discussion after $18 \ln n$ rounds we can lower bound the probability of pulling a node with any color $i \in I$ by $\frac{\frac{n}{2k}-1}{n} \approx \frac{1}{2k}$. By Lemma 3.1, for all $v \in V$, among additional $48 \ln n + 12k$ rounds, at least $2k$ are collision-free rounds for v. During these additional $2k$ collision-free rounds for v, we obtain $Pr[i \notin M_v(r)] \leq (1 - \frac{1}{2k})^{2k} \leq e^{-1}$, therefore $Pr[i \in M_v(r)] > \frac{1}{2}$. Note that, according to Lemma 3.2 during those additional rounds all colored nodes have obtained their maximal age w.h.p., hence, the effective cases of our exchange policy 3.1 are the last two. □

Theorem 3.4 *(Phase 3) The protocol spreads all messages among all nodes in $O(k + \ln n)$ rounds w.h.p.*

Proof. Denote $t = 24k + 192 \ln n$ and $r_0 = 66 \ln n + 12k$. According to Lemma 3.1, for all $v \in V$ there are $t/6$ collision-free rounds for v between round r_0 and $r_0 + t$. Let $y_v(r)$ be a random variable indicating if v succeeds in pulling a new message in a collision-free round r. According to Lemma 3.3, $\forall u \in V$ the independent probability that in any collision-free round $r > r_0$ u obtains any specific message $i \in I$ is at least $\frac{1}{2}$. Let Y_v denote the sum of $y_v(r)$ over the $t/6$ collision-free rounds r for v. We obtain that $E[Y_v] \geq t/12$. Applying again a Chernoff bound we obtain:

$$\begin{aligned} Pr[|M_v(t)| < k] &\leq Pr[Y_v < k] \\ &= Pr[Y_v < (1-\delta)\mu] \\ &\leq \exp\left(\frac{-\mu\delta^2}{2}\right) \\ &< \frac{1}{n^2} \end{aligned}$$

where $\mu = 2k + 16 \ln n$, $\delta = \frac{1}{2}$. From this, the high probability result follows by a union bound. □

Theorem 3.5 *Let m denote the size of the initial messages input to any processor $i \in I$. The total number of communication bits employed by the protocol is*

$$knm + O(n^2 \ln n + kn^2) \ .$$

Proof. Since the protocol uses the pull model, each node is a recipient of a single message in each round. Since each node updates its message set every round, there are no redundant message transmissions executed by the protocol. The communication costs are therefore comprised of the following two parts:

1. knm communication associated with data transferred.
2. $O(n^2 \ln n + kn^2)$ communication associated with pull requests.

□

4 Mass Conservation

Thus far, our protocol relied on prior knowledge of the number of processor nodes, n, and the initial message set size, k. This served both in the aging mechanism (e.g. $\log \frac{n}{2k}$) , and in order for a node to locally terminate the protocol (i.e., to stop pulling when it had obtained k messages).

In this section, we present a simple bootstrap protocol through which nodes learn the message set size, k given the number of processor nodes, n. The proposed procedure is based on the *push-sum* protocol introduced in [KDG03], additionally [BGPS05, DSW06] undertake an in-depth study of the design and analysis for averaging in an arbitrarily connected network of nodes, however since our system model assumes full connectivity the simplified version presented in [KDG03] suffice. In the paper, the authors investigate the problem of computing aggregates (e.g. sums and averages) with gossip-style protocols. The protocol is fairly simple, in each round r, each node $i \in [n]$ maintains a sum $s_{r,i}$ initialized to arbitrary value $s_{0,i} = x_i$, and a weight $w_{r,i}$, initialized to $w_{0,i} = 1$. In each subsequent round $r > 0$, each node i computes the pair $\{s_{r,i}, w_{r,i}\}$, as the sum of all the values and weights it received during round $r - 1$ respectively. Then the node chooses a target uniformly at random and sends the pair $\{\frac{1}{2}s_{r,i}, \frac{1}{2}w_{r,i}\}$ to the target node and to itself. The protocol analysis shows that the values $\frac{s_{r,i}}{w_{r,i}}$ at all nodes i converge exponentially fast to the average, $\frac{\sum_{i=1}^{n} x_i}{n}$. More concretely, with probability at least $1 - \frac{1}{\delta}$, the relative error in the approximation of the average drops within ε, in at most $O(\log n + log\frac{1}{\varepsilon} + \log\frac{1}{\delta})$ rounds, where in each round each node sends a single message. In our bootstrap protocol we are interested in computing the size of $|I| = k$ instead of average, then we only need to apply small change: $\forall v \in I s_{0,v} = 1$ otherwise $\forall v \notin I s_{0,v} = 0$ and only one node starts with weight 1 and the rest with 0. Once this protocol completes, every node has obtained the size of k w.h.p. and can carry the multiple message mongering protocol.

Gossip algorithms for aggregation of information have recently received significant attention for various types of network applications because of their simplicity and inherent robustness. To name only few [KDG03, BGPS05, DSW06, JYCX05] these algorithms are very natural and simple, yet the proof of these algorithms guarantee is non-trivial and relies crucially on a useful property of *mass conservation*. For instance, in [KDG03] the average of all sums $s_{r,i}$ is always the correct average, and the sum of all weights $w_{r,i}$ is always n. Surprisingly enough, our aging mechanism also maintains *mass conservation* of each colored group, in each round r, $\sum_{v \in C(i,r)} 2^{-a_v} = 1$. This way the protocol limits the scope of primary messages dissemination to $n/2k$ nodes, in order to overcome the skewness of spreading k different messages using randomized gossip (e.i., 'rare' chunks). As a result, our protocol result in an asymptotically optimal time algorithm w.h.p.

5 Reducing the Message Payload

In the protocol above, u sends at every round to its gossip partner the vector $M_u(r)$. This incurs an overhead of n bits associated with each and every communication. When n is large, especially if $n > m$, this cost may become prohibitive. Interestingly, a more careful investigation of the protocol's analysis allows us to reduce this overhead, albeit at increased data transfer costs. In this section, we describe a variant of our protocol, which uses only $\log n$ bits payload associated with each communication.

The reduced-payload protocol employs a similar pull-based framework, where in each round $r > 0$, every node u chooses a gossip partner v uniformly at random. Node u sends v a message index $i \in I \setminus M_u(r)$ and an indication whether it has color or not. Node v acts as follows:

1. When v has color c_v, u does not have a color, and c_v's age, a_v, satisfies $a_v < \log\frac{n}{2k}$, v increments c_v's age to $a_v + 1$, and sends $\langle c_v, a_v \rangle$ to u.

2. Otherwise, if $r < 18 \ln n + 2k$ then v sends its color message c_v.

3. Otherwise, if $i \in M_v(r)$ then v sends the message i.

When u gets the response from v, it stores it locally. If this is a color message with an age, then u becomes colored with color c_v and stores its age along with it.

The time analysis in Section 3.2 above carries over to the reduced-payload protocol. Hence, we do not repeat the proofs, and only briefly comment on the relationship here.

By the end of the first phase, according to Lemma 3.2, every message color set includes $\frac{n}{2k} - 1$ nodes w.h.p. In order to implement this phase the protocol relies on a single bit indicating whether the pull request came from a colored node or uncolored one.

In the second phase, Lemma 3.3, the analysis rests on the fact that a colored node always prefers to propagate its own message color. Hence, there is no need for message information exchange between communicating nodes.

The third phase is addressed in Theorem 3.4. The proof relies on the fact (Lemma 3.3) that the probability that a node has acquired any particular message is at least $\frac{1}{2}$. Accordingly, the only information required for a successful message exchange to occur with probability $\frac{1}{2}$ is an index of any selected missing message. We remark that although the reduced-payload protocol has the same asymptotic completion time as the first protocol version, the first protocol clearly dominates it. That is, whenever a node obtains a new message in the reduced-payload protocol, it also obtains it in the first version. However, the reverse may not hold.

The message complexity of the reduced-payload protocol is given by the following theorem.

Theorem 5.1 *Let m denote the size of the initial messages input to any processor $i \in I$. The total number of communication bits employed by the protocol is*

$$O((k + \ln n)n(m + \log n)) \ .$$

Proof. The protocol works in $O(k + \ln n)$ rounds. In each round, every node sends at most m bits of message data, as well as $O(\log n)$ control data. □

6 Related Work

In early stages, gossip protocols investigation has mainly focused on single source gossip. Notably, Demers et al. [DGH$^+$87] performed a detailed study of epidemic algorithms, in which a message (update) is initially known at a single processor and must be diffused to all processors with minimal traffic overhead. One of the algorithms they studied, called *anti-entropy* and apparently initially proposed in [BLNS82], was adopted in Xerox's Clearinghouse project (see [DGH$^+$87]) and MUSE (for USENET News propagation) [LOM94]. Similar ideas also proposed as message loss detection and recovery techniques in multicast protocols [Dee89, BHOO$^+$99, SS00, EGH$^+$03, TM04]. The protocols are generally composed of two sub-protocols. The first is an unreliable hierarchical broadcast that makes a best-effort attempt to efficiently deliver each message to its destinations. The second is an anti-entropy protocol that operates in a series of unsynchronized rounds. During each round, the first phase detects message loss; the second phase corrects such losses and runs only if needed. This approach was proved to scale better than the traditional deterministic techniques by using simulations and non formal analysis.

Recent work [DM04] considers the problem of multiple rumors mongering only for the case that the initial set size, k, is $\Theta(n)$. It proposes a scheme based on random network coding. The protocol achieves an asymptotically optimal running time of $O(n)$ rounds w.h.p. Compared with our approach, their solution does not address the case where $k = o(n)$. Additionally, it has a prohibitive computation cost of creating linear combination in each round and a payload of $n \log n$ bits associated with each message. A conjecture is made in [DM04] that any store and forward protocol, in other words, a protocol that does not allow the messages to be manipulated, can do no better than $O(k \ln n)$. Our protocol serves to disprove this conjecture. It is a store and forward that achieves an asymptotically optimal running time of $O(k + \ln n)$ rounds w.h.p. for any initial set size k (in particular, $O(n)$ rounds when $k = O(n)$), and sends the original messages un-altered with payload of $\log n$ bits associated with each message. In addition to simplicity, our approach makes parts of the data available for the downloading participant as soon as they arrive. This can be quite useful in the case of large images and even in some videocasts. With network coding, a participant has to retrieve most, or all of the chunks, in order to decode any of the pieces. Thus, no content is available until the final stages of the protocol. In the case of large images, it is beneficial to be able to view partial data as soon as it is available.

Karp et al. [KSSV00] show that any address-oblivious algorithm (i.e., an algorithm that does not use the initiator's address in determining communication partners) needs to establish $\Omega(n \ln \ln n)$ connections for each rumor regardless of the number of rounds. This result helps emphasize the dramatic effect of interleaving on the overall performance in multi-message gossip: The total connections employed by our scheme for k messages gossip is $O(n(k + \ln n))$. When k is large, this is better than repeating k times $O(n \ln \ln n)$ connections. It is left open to find whether techniques from [KSSV00] may be employed to reduce our connection costs to $O(n(k + \ln \ln n))$.

Gossip algorithms that compute certain aggregate functions of information from the network are studied in [KDG03], and extended to arbitrary network topologies in [BGPS05]. For complete graphs, convergence occurs in $O(\log n)$ rounds w.h.p. and with message complexity of $O(n \log^2 n)$.

7 Conclusions

As internets become increasingly used for the wide-scale broadcast of information, the ability to send packets to large fractions of the internet at near-optimal cost may be the vital step that will allow the internet to replace traditional broadcast media. For large multicast groups, there are substantial inefficiencies that result from using a multicast tree to send messages to many recipients, both from the standpoint of the sender, and from the standpoint of wasted aggregate bandwidth.

Our work contributes to the understanding of collaborative content dissemination. The advantages of collaborative delivery lie in its robustness, in its total bandwidth scaling, and the improved overall completion time. We investigate the power of randomized gossip for shared download. We present a simple and clean solution with promising behavior. It achieves asymptotically optimal delivery time while incurring minimal payload overhead and connection set-up costs.

References

[BGPS05] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Gossip algorithms: Design, analysis and applications. In *Proceedings of IEEE INFOCOM*, 2005.

[BHOO$^+$99] K. P. Birman, M. Hayden, Z. Xiao O. Ozkasap, M. Budio, and Y. Minsky. Bimodal multicast. *ACM Transactions on Computer Systems*, 17(2):41–88, 1999.

[BLNS82] A. D. Birrell, R. Levin, R. M. Needham, and M. D. Schroeder. Grapevine, an exercise in distributed computing. *Communications of the ACM*, 25(4):260–274, 1982.

[Dee89] S. E. Deering. Host extensions for ip multicasting. Technical Report RFC 1112, SRI Network Information Center, August 1989.

[DGH^{+}87] Alan Demers, Dan Greene, Carl Hauser, Wes Irish, John Larson, Scott Shenkcr, Howard Sturgis, Dan Swinehart, and Doug Terry. Epidemic algorithms for replicated database maintenance. In *PODC '87: Proceedings of the sixth annual ACM Symposium on Principles of distributed computing*, pages 1–12, New York, NY, USA, 1987. ACM Press.

[DM04] S. Deb and M. Medard. Algebraic gossip: A network coding approach for optimal multiple rumor mongering. In *Allerton Conference on Communication, Control, and Computing*, 2004.

[DSW06] A. G. Dimakis, A. D. Sarwate, and M. Wainwright. Geographic gossip : Efficient aggregation for sensor networks. In *5th International Symposium on Information Processing in Sensor Networks (IPSN 2006)*, Nashville, TN, April 2006.

[EGH^{+}03] P. Th. Eugster, R. Guerraoui, S. B. Handurukande, P. Kouznetsov, and A.-M. Kermarrec. Lightweight probabilistic broadcast. *ACM Trans. Comput. Syst.*, 21(4):341–374, 2003.

[GHK04] R. Guerraoui, S. B. Handurukande, and A.-M. Kermarrec. GosSkip: a Gossip-based Structured Overlay Network for Efficient Content-based Filtering. Technical report, 2004.

[GKM03] A. J. Ganesh, Anne-Marie Kermarrec, and Laurent Massouli. Peer-to-peer membership management for gossip-based protocols. *IEEE Transactions on Computers*, 52(2), February 2003.

[GR05] Christos Gkantsidis and Pablo Rodriguez. Network coding for large scale content distribution. In *Proceedings of Infocom 2005*. IEEE, IEEE, March 2005.

[HHL02] Z. Haas, J.Y. Halpern, and L. Li. Gossip-based ad hoc routing. In *Proceedings of IEEE INFOCOM*, June 2002.

[JLC05] Kamal Jain, László Lovász, and Philip A. Chou. Building scalable and robust peer-to-peer overlay networks for broadcasting using network coding. In *PODC '05: Proceedings of the twenty-fourth annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing*, pages 51–59, New York, NY, USA, 2005. ACM Press.

[JYCX05] G. Pandurangan J. Y. Chen and D. Xu. Robust computation of aggregates in wireless sensor networks: distributed randomized algorithms and analysis. In *4th International Symposium on Information Processing in Sensor Networks (IPSN)*, Los Angeles, CA, April 2005.

[KB05] Yoram Kulbak and Danny Bickson. The emule protocol speficiation. Technical Report Leibniz Center TR-2005-03, School of Computer Science and Engineering, The Hebrew University, 2005.

[KDG03] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *FOCS '03: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, page 482, Washington, DC, USA, 2003. IEEE Computer Society.

[KKD04] David Kempe, Jon Kleinberg, and Alan Demers. Spatial gossip and resource location protocols. *J. ACM*, 51(6):943–967, 2004.

[KSSV00] R. Karp, C. Schindelhauer, S. Shenker, and B. Vöcking. Randomized rumor spreading. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 565, Washington, DC, USA, 2000.

[LOM94] K. Lidl, J. Osborne, and J. Malcome. Drinking from the firehose: Multicast usenet news. In *Proceedings of the Usenix Winter Conference*, pages 33–45, January 1994.

[SHL88] S.T. Hedetniemi S.M. Hedetniemi and A.L. Liestman. A survey of gossiping and broadcasting in communication networks. *Networks*, 18(1):319–349, 1988.

[SS00] Qixiang Sun and Daniel C. Sturman. A gossip-based reliable multicast for large-scale high-throughput applications. In *DSN '00: Proceedings of the 2000 International Conference on Dependable Systems and Networks (formerly FTCS-30 and DCCA-8)*, page 347, Washington, DC, USA, 2000. IEEE Computer Society.

[TM04] Soontaree Tanaraksiritavorn and Shivakant Mishra. Evaluation of gossip to build scalable and reliable multicast protocols. *Perform. Eval.*, 58(2+3):189–214, 2004.

[VGvS05] Spyros Voulgaris, Daniela Gavidia, and Maarten van Steen. Cyclon: Inexpensive membership management for unstructured p2p overlays. *J. Network Syst. Manage.*, 13(2), 2005.

[vR00] R. van Renesse. Scalable and secure resource location. In *Proceedings of IEEE Hawaii International Conference on System Sciences*, January 2000.

[vRMH98] R. van Renesse, Y. Minsky, and M. Hayden. A gossip-style failure detection service. In *Proceedings of Middleware*, 1998.

APPENDIX

We detail here the proof outlined in Section 3.2 of Lemma 3.2.

Lemma .1 *For all colors $i \in I$ and all rounds r, at most $\frac{n}{2k} - 1$ nodes are colored with color i.*

Proof. According to our exchange policy 3.1(1) a colored message age, a_i, cannot exceed $\log \frac{n}{2k}$. Consequently, $\forall i \in I$, the number of colored nodes with the same color $c(i, r)$, can be bounded by: $\sum_{a_i=0}^{\log \frac{n}{2k} - 1} 2^{a_i} = 2^{\log \frac{n}{2k}} - 1 = \frac{n}{2k} - 1$. □

Lemma .2 *For all colors $i \in I$ and all rounds $r > 18 \ln n$, at least $\frac{n}{2k} - 1$ nodes are colored with color i w.h.p.*

Proof. Let $p(r)$ denote the probability that a colored node is pulled by an uncolored one during round r. Since half the nodes are uncolored at any time by Lemma .1, $p(r)$ is always lower bounded by $\frac{1}{3}$.

$$p(r) \geq 1 - (1 - \frac{1}{n})^{\frac{n}{2}} \geq 1 - e^{-\frac{1}{2}} > \frac{1}{3}$$

Fix any node v, and as usual, denote by $c_v \in I$ its color. Denote by $a_{v,0}$ the initial age of the color message c_v when obtained by v, $0 \leq a_{v,0} \leq \log \frac{n}{2k}$. Given the round in which the colored node was first colored, $r_0 \leq r$, and its initial age, $a_{v,0}$, the age $a_{v,r}$ of the color message c_v at node v in the r'th round is lower bounded by the following binomially distributed random variable in stochastic ordering sense.

$$a_{v,r} \succ Bin(r - r_0, \frac{1}{3}) + a_{v,0}$$

Let $r_{v,0}$ be a random variable that indicates the round in which v obtained its color. We first show that for any given $a_{v,0}$, $E[r_{v,0} \mid a_{v,0}] \leq 3a_{v,0}$.

The proof is by induction on the initial age $a_{v,0}$. In the initial step when $a_{v,0} = 1$, the color message was the first to be pulled directly from the node $i \in I$ where $i = c_v$(note that here we use $i \in I$ to uniquely indicate the origin node in I that initiated this message). Therefore, $r_{v,0}$ is a geometric random variable with probability greater than $\frac{1}{3}$, which indicates the sequence of rounds until node c_v is pulled, with expectation $E[r_{v,0} \mid a_{v_0} = 1] \leq 3$.

Assume that, for an arbitrary k, $a_{v,0} \leq k$ is also true, we now derive $a_{v,0} = k + 1$ from this assumption. Consider the node u from which v obtains its color $c_v = c_u$ during round $r_{v,0}$ with an initial age $a_{v,0} = k + 1$. Since the initial age of u satisfies $a_{u,0} \leq k$, the inductive assumption holds, $E[r_{u,0} \mid a_{u,0} \leq k] \leq 3a_{u,0}$. Furthermore, u has succeeded in increasing its age $a_{v,0} - a_{u,0}$ times. The number of rounds $r_{v,0} - r_{u,0}$ for u to succeed $a_{v,0} - a_{u,0}$ times is a negative binomial distribution random variable, $E[r_{v,0} - r_{u,0} \mid a_{v,0} - a_{u,0}] \leq 3(a_{v,0} - a_{u,0})$. Luckily, we can get rid of the dependency on $a_{u,0}$ by considering the following expected sum:

$$\begin{aligned}
E[r_{v,0} \mid a_{u,0}, a_{v,0}] &= E[r_{u,0} + (r_{v,0} - r_{u,0}) \mid a_{u,0}, a_{v,0}] \\
&\leq 3a_{u,0} + 3(a_{v,0} - a_{u,0}) \\
&= 3a_{v,0}
\end{aligned}$$

Since $a_{v,r} \succ Bin(r - r_0, \frac{1}{3}) + a_{v,0}$, we can define a conditional distribution

$$\psi(a_{v,0}) = E[a_{v,r} \mid a_{v,0}] \geq a_{v,0} + (r - E[r_{v,0} \mid a_{v,0}])\frac{1}{3}$$

We can compute the following expectation:

$$\begin{aligned}
E[a_{v,r}] &= E[\psi(a_{v,0})] \\
&\geq E[a_{v,0} + (r - E[r_{v,0} \mid a_{v,0}])\frac{1}{3}] \\
&\geq E[a_{v,0} + (r - 3a_{v,0})\frac{1}{3}] \\
&= \frac{r}{3}
\end{aligned}$$

Conclude that all colored nodes have the same age expectation:

$$\forall a_{v,0} \in [\log \frac{n}{2k}] \;\; E[a_{v,r}] \geq \frac{r}{3}.$$

Since the pull events at different rounds are independent, we may apply Chernoff bounds to obtain:

$$\begin{aligned} Pr[a_{v,r} \leq \ln n] &= Pr[a_{v,r} \leq (1-\delta)\mu] \\ &\leq \exp\,(\frac{-\mu\delta^2}{2}) \\ &< n^{-2} \end{aligned}$$

Where $r = 18 \ln n$, hence $rp = \mu \geq 6 \ln n$, and $\delta = 5/6$.

The lower bound on the age of all nodes v with color $c_v = i \in I$ indicates that the color-set satisfies $c(i, 18 \ln n) \geq \frac{n}{2k} - 1$. By a union bound, the probability that all colors $i \in I$ satisfy $c(i, 18 \ln n) = \frac{n}{2k} - 1$ is at least $1 - 1/n$, which completes the proof. □