Censorship Resistant Peer-to-Peer Content Addressable Networks

Amos Fiat[*] Jared Saia[†]

Thomas Hobbes, Rene Descartes, Francis Bacon, Benedict Spinoza, John Milton, John Locke, Daniel Defoe, David Hume, Jean-Jacques Rousseau, Blaise Pascal, Immanual Kant, Giovanni Casanova, John Stuart Mill, Emile Zola, Jean-Paul Sartre, Victor Hugo, Honore de Balzac, A. Dumas pere, A. Dumas fil, Gustave Flaubert, Rabelais, Montaigne, La Fontaine, Voltaire, Denis Diderot, Pierre Larousse, Anatole France

Partial List of authors in Index Librorum Prohibitorum *(Index of Prohibited Books) from the Roman Office of the Inquisition, 1559–1966.*

Abstract

We present a censorship resistant peer-to-peer Content Addressable Network for accessing n data items in a network of n nodes. Each search for a data item in the network takes $O(\log n)$ time and requires at most $O(\log^2 n)$ messages. Our network is censorship resistant in the sense that even after adversarial removal of an arbitrarily large constant fraction of the nodes in the network, all but an arbitrarily small fraction of the remaining nodes can obtain all but an arbitrarily small fraction of the original data items. The network can be created in a fully distributed fashion. It requires only $O(\log n)$ memory in each node. We also give a variant of our scheme that has the property that it is highly spam resistant: an adversary can take over complete control of a constant fraction of the nodes in the network and yet will still be unable to generate spam.

1 Introduction

Web content is under attack by state and corporate efforts to censor it, for political and commercial reasons ([7, 18, 16]).

Peer-to-peer networks are considered more robust against censorship than standard web servers ([19]). However, while it is true that many suggested peer-to-peer architectures are fairly robust against random faults, the censors can attack carefully chosen weak points in the system. For example, the Napster ([28]) file sharing system has been effectively dismembered by legal attacks on the central server. Additionally, the Gnutella ([27]) file sharing system, while specifically designed to avoid the vulnerability of a central server, is highly vulnerable to attack by removing a very small number of carefully chosen nodes ([24]).

A significant performance problem with Gnutella is that it requires up to $O(n)$ messages for a search (search is performed via a general broadcast), this has effectively limited the size of Gnutella networks to about 1000 nodes ([6]).

A more principled approach than the centralized approach taken by Napster or the broadcast search mechanism of Gnutella is the use of a content addressable network ([23]). A content addressable network is defined as a distributed, scalable, indexing scheme for peer-to-peer networks. Plaxton, Rajaram and Richa ([22]) give a scheme to implement a context addressable network (prior to its definition) in a web cache environment. The content addressable network scheme of ([22]), as modified in ([29]), has been used in implementations such as Oceanstore ([11]). Subsequent content addressable networks have been suggested in ([23, 26, 29]). These papers seek to improve upon ([22]) in adding fault resistance and load balancing mechanisms. However, while these schemes are robust to a large number of random faults, none of these schemes are robust against adversarial faults.

1.1 Resistance to Adversarial Node Deletion

We present a content addressable network with n nodes used to store n distinct data items[1]. As far as we know this is the first such scheme of its kind. The scheme is robust to adversarial deletion of up to[2] half of the nodes in the network and has the following properties:

[*]Department of Computer Science, Tel Aviv University, work done while on Sabbatical at University of Washington, Seattle. email: fiat@math.tau.ac.il

[†]Department of Computer Science, University of Washington, Seattle. email: saia@cs.washington.edu.

[1]For simplicity, we've assumed that the number of items and the number of nodes is equal. However, for any n nodes and $m \geq n$ data items, our scheme will work, where the search time remains $O(\log n)$, the number of messages remains $O(\log^2 n)$, and the storage requirements are $O(\log n \times m/n)$ per node.

[2]For simplicity, we give the proofs with this constant equal to 1/2. However we can easily modify the scheme to work for any constant less than 1. This would change the constants involved in storage, search time, and messages sent, by a constant factor.

1. With high probability, all but an arbitrarily small fraction of the nodes can find all but an arbitrarily small fraction of the data items.

2. Search takes (parallel) time $O(\log n)$.

3. Search requires $O(\log^2 n)$ messages in total.

4. Every node requires $O(\log n)$ storage.

As stated above, in the context of state or corporate attempts at censorship, it seems reasonable to consider *adversarial* attacks rather than random deletion. Our scheme is a content addressable network that is robust against adversarial deletion.

We remark that such a network is clearly resilient to having up to 1/2 of the nodes removed at random, (in actuality, its random removal resiliency is much better). We further remark that if nodes come up and down over time, our network will continue to operate as required so long as at least $n/2$ of the nodes are alive.

1.2 Spam Resistance Another problem with peer-to-peer networks has been that of spamming ([5, 6]). Because the data items reside in the nodes of the network, and pass through nodes while in transit, it is possible for nodes to invent alternative data and pass it on as though it was the sought after data item.

We now describe a spam resistant variant of our content addressable network. To the best of our knowledge this is the first such scheme of its kind. As before, assume n nodes used to store n distinct data items. The adversary may choose up to some constant $c < 1/2$ fraction of the nodes in the network. These nodes under adversary control may be deleted, or they may collude and transmit arbitrary false versions of the data item, nonetheless:

1. With high probability, all but an arbitrarily small fraction of the nodes will be able to obtain all but an arbitrarily small fraction of the *true* data items. To clarify this point, the search will *not* result in multiple items, one of which is the correct item. The search will result in one unequivocal true item.

2. Search takes (parallel) time $O(\log n)$.

3. Search requires $O(\log^3 n)$ messages in total.

4. Every node requires $O(\log^2 n)$ storage.

The rest of our paper is structured as follows. We review related work in Section 2. We give the algorithm for creation of our robust content addressable network, the search mechanism, and properties of the content addressable network in Section 3. The proof of our main theorem, Theorem 3.1, is given in Section 4. In Section 5 we sketch the modifications required in the algorithms and the proofs to obtain spam resistance, the main theorem with regard to spam resistant content addressable networks is Theorem 5.1. We conclude and give directions for future work in Section 6. Acknowledgements are in section 7.

2 Related Work

2.1 Peer-to-peer Networks (not Content Addressable) Peer-to-peer networks are a relatively recent and quickly growing phenomena. The average number of Gnutella users in any given day is no less than $10,000$ and may range as high as $30,000$ ([6]). Napster software has been downloaded by 50 million users ([23]).

Pandurangam, Raghavan, and Upfal ([20]) address the problem of maintaining a connected network under a probabilistic model of node arrival and departure. They do not deal with the question of searching within the network. They give a protocol which maintains a network on n nodes with diameter $O(\log n)$. The protocol requires constant memory per node and a central hub with constant memory with which all nodes can connect.

Experimental measurements of a connected component of the real Gnutella network have been studied ([24]), and it has been found to still contain a large connected component even with a 1/3 fraction of random node deletions.

2.2 Content Addressable Networks — Random Faults There are several papers that address the problem of creating a content addressable network. As mentioned above, Plaxton, Rajaram and Richa ([22]) give a context addressable network for web caching. Search time and the total number of messages is $O(\log n)$, and storage requirements are $O(\log n)$ per node.

Tapestry ([29]) is an extension to the ([22]) mechanism, designed to be robust against faults. It is used in the Oceanstore ([11]) system. Experimental evidence is supplied that Tapestry is robust against random faults.

Ratnasamy *et. al.* ([23]) describe a system called CAN which has the topology of a d-dimensional torus. As a function of d, storage requirements are $O(d)$ per node, whereas search time and the total number of messages is $O(dn^{1/d})$. There is experimental evidence that CAN is robust to random faults.

Finally, Stoica *et. al.* introduce yet another content addressable network, Chord ([26]), which, like ([20]) and ([29]), requires $O(\log n)$ memory per node and $O(\log n)$ search time. Chord is *provably* robust to a constant fraction of random node failures.

2.3 Faults on Networks

2.3.1 Random Faults There is a large body of work on node and edge faults that occur independently at random in a general network. Håstad, Leighton and Newman ([9]) address the problem of routing when there are node and edge faults on the hypercube which occur independently at random with some probability $p < 1$. They give a $O(\log n)$ step routing algorithm that ensures the delivery of messages with high probability even when a constant fraction of the nodes and edges have failed. They also show that a faulty hypercube can emulate a fault-free hypercube with only constant slowdown.

Karlin, Nelson and Tamaki ([10]) explore the fault tolerance of the butterfly network against edge faults that occur independently at random with probability p. They show that there is a critical probability p^* such that if p is less than p^*, the faulted butterfly almost surely contains a linear-sized component and that if p is greater than p^*, the faulted butterfly does not contain a linear sized component.

Leighton, Maggs and Sitamaran ([12]) show that a butterfly network whose nodes fail with some constant probability p can emulate a fault-free network of the same size with a slowdown of $2^{O(\log^* n)}$.

2.3.2 Adversarial Faults It is well known that many common network topologies are not resistant to a linear number of adversarial faults. With a linear number of faults, the hypercube can be fragmented into components all of which have size no more than $O(n/\sqrt{\log n})$ ([9]). The best known lower bound on the number of adversarial faults a hypercube can tolerate and still be able to emulate a fault free hypercube of the same size is $O(\log n)$ ([9]).

Leighton, Maggs and Sitamaran ([12]) analyze the fault tolerance of several bounded degree networks. One of their results is that any n node butterfly network containing $n^{1-\epsilon}$ (for any constant $\epsilon > 0$) faults can emulate a fault free butterfly network of the same size with only constant slowdown. The same result is given for the shuffle-exchange network.

2.4 Other Related Work

One attempt at censorship resistant web publishing is the Publius system ([15]), while this system has many desirable properties, it is not a peer-to-peer network. Publius makes use of many cryptographic elements and uses Shamir's threshold secret sharing scheme ([25]) to split the shares amongst many servers. When viewed as a peer-to-peer network, with n nodes and n data items, to be resistant to $n/2$ adversarial node removals, Publius requires $\Omega(n)$ storage per node and $\Omega(n)$ search time per query.

Alon et al. ([1]) give a method which safely stores a document in a decentralized storage setting where up to half the storage devices may be faulty. The application context of their work is a storage system consisting of a set of servers and a set of clients where each client can communicate with all the servers. Their scheme involves distributing specially encoded pieces of the document to all the servers in the network.

Aumann and Bender ([3]) consider tolerance of pointer-based data structures to worse case memory failures. They present fault tolerant variants of stacks, lists and trees. They give a fault tolerant tree with the property that if r adversarial faults occur, no more than $O(r)$ of the data in the tree is lost. This fault tolerant tree is based on the use of expander graphs.

Quorum systems ([8, 13, 14]) are an efficient, robust way to read and write to a variable which is shared among n servers. Many of these systems are resistant up to some number $b < n/4$ of Byzantine faults. The key idea in such systems is to create subsets of the servers called *quorums* in such a way that any two quorums contain at least $2b + 1$ servers in common. A client that wants to write to the shared variable will broadcast the new value to all servers in some quorum. A client that wants to read the variable will get values from all members in some quorum and will keep only that value which has the most recent time stamp and is returned by at least $b + 1$ servers. For quorum systems that are resistant to $\theta(n)$ faults the load on the servers can be high. In particular, $\theta(n)$ servers will be involved in a constant fraction of the queries.

Recently Malkhi *et. al.* [14] have introduced a probabilistic quorum system. This new system relaxes the constraint that there must be $2b + 1$ servers shared between any two quoroms and remains resistant to Byzantine faults only with high probability. The load on servers in the probabilistic system is less than the load in the deterministic system. Nonetheless, for a probabilistic quorum system which is resistant to $\theta(n)$ faults, there still will be at least one server involved in a constant fraction of the queries.

3 The Content Addressable Network

We now state our mechanism for providing indexing of n data items by n nodes in a network that is robust to removal of any $n/2$ of the nodes. We make use of a butterfly network of depth $\log n - \log\log n$, we call the nodes of the butterfly network *supernodes* (see Figure 1). Every supernode is associated with a set of nodes. We call a supernode at the topmost level of the butterfly a top supernode, one at the bottommost level of the network a bottom supernode and one at neither the

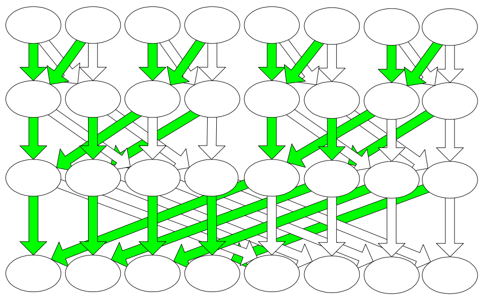

Figure 1: The butterfly network of supernodes.

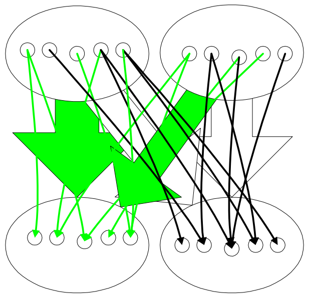

Figure 2: The expander graphs between supernodes.

topmost or bottommost level a middle supernode.

To construct the network we do the following:

- We choose an error parameter $\epsilon > 0$, and as a function of ϵ we determine constants C, B, T, D, α and β. (See Theorem 3.1).
- Every node chooses at random C top supernodes, C bottom supernodes and $C \log n$ middle supernodes to which it will belong.
- Between two sets of nodes associated with two supernodes connected in the butterfly network, we choose a random constant degree expander graph of degree D (see Figure 2). (We do this only if both sets of nodes are of size at least $\alpha C \ln n$ and no more than $\beta C \ln n$.)
- We also map the n data items to the $n/\log n$ bottom supernodes in the butterfly. Every one of the n data items is hashed to B random bottom supernodes. (Typically, we would not hash the entire data item but only it's title, *e.g.*, "Singing in the Rain"). [3]
- The data item is stored in all the component nodes of all the (bottom) supernodes to which it has been hashed (if any bottom supernode has more than $\beta B \ln n$ data items hashed to it, it drops out of the network.)
- In addition, every one of the nodes chooses T top supernodes of the butterfly and points to all component nodes of these supernodes.

[3]We use the random oracle model ([4]) for this hash function, it would have sufficed to have a weaker assumption such as that the hash function is expansive.

To perform a search for a data item, starting from node v, we do the following:

```
1. Take the hash of the data item and interpret it as a sequence of indices $i_1, i_2, \ldots, i_B$, $0 \le i_\ell \le n/\log n$.
2. Let $t_1, t_2, \ldots, t_T$ be the top supernodes to which $v$ points.
3. Repeat in parallel for all values of $k$ between 1 and $T$:
   (a) Let $\ell = 1$.
   (b) Repeat until successful or until $\ell > B$:
       i. Follow the path from $t_k$ to the supernode at the bottom level whose index is $i_\ell$:
          • Transmit the query to all of the nodes in $t_k$. Let $W$ be the set of all such nodes.
          • Repeat until a bottom supernode is reached:
            – The nodes in $W$ transmit the query to all of their neighbors along the (unique) butterfly path to $i_\ell$, Let $W$ be this new set of nodes.
          • When the bottom supernode is reached, fetch the content from whatever node has been reached.
          • The content, if found, is transmitted back along the same path as the query was transmitted downwards.
       ii. Increment $\ell$.
```

3.1 Properties of the Content Addressable Network

Following is the main theorem which we will prove in Section 4.

THEOREM 3.1. *For all $\epsilon > 0$, there exist constants $k_1(\epsilon)$, $k_2(\epsilon)$, $k_3(\epsilon)$ which depend only on ϵ such that*

- *Every node requires $k_1(\epsilon) \log n$ memory.*
- *Search for a data item takes no more than $k_2(\epsilon) \log n$ time.*
- *Search for a data item requires no more than $k_3(\epsilon) \log^2 n$ messages.*
- *All but ϵn nodes can reach all but ϵn data items.*

3.2 Some Comments

1. **Distributed creation of the content addressable network**

   We note that our Content Addressable Memory can be created in a fully distributed fashion with n broadcasts or transmission of n^2 messages in total and assuming $O(\log n)$ memory per node. We briefly sketch the protocol that a particular node will follow to do this. The node first randomly chooses the supernodes to which it belongs. Let S be the set of supernodes which neighbors supernodes to which the node belongs. For each $s \in S$, the node chooses a set N_s of D random numbers between 1 and $\beta C \ln n$. The node then broadcasts a message to all other nodes which contains the identifiers of the supernodes to which the node belongs.

   Next, the node will receive messages from all other nodes giving the supernodes to which they belong. For every $s \in S$, the node will link to the i-th node that belongs to s from which it receives a message if and only if $i \in N_s$.

   If for some supernode to which the node belongs, the node receives less than $\alpha C \ln n$ or greater than $\beta C \ln n$ messages from other nodes in that supernode, the node removes all out-going connections associated with that supernode. Similarly, if for some supernode in S, the node receives less than $\alpha C \ln n$ or greater than $\beta C \ln n$ messages from other nodes in that supernode, the node removes all out-going connections to that neighboring supernode. Connections to the top supernodes and storage of data items can be handled in a similar manner.

2. **Insertion of a New Data Item**

   One can insert a new data item simply by performing a search, and sending the data item along with the search. The data item will be stored at the nodes of the bottommost supernodes in the search. We remark that such an insertion may fail with some small constant probability.

3. **Insertion of a New Node**

   Our network does not have an explicit mechanism for node insertion. It does seem that one could insert the node by having the node choose at random appropriate supernodes and then forging the required random connections with the nodes that belong to neighboring supernodes. The technical difficulty with proving results about this insertion process is that not all live nodes in these neighboring supernodes may be reachable and thus the probability distributions become skewed.

   We note though that a new node can simply copy the links to top supernodes of some other node already in the network and will thus very likely be able to access almost all of the data items. This insertion takes $O(\log n)$ time. Of course the new node will not increase the resiliency of the network if it inserts itself in this way. We assume that a full reorganization of the network is scheduled whenever sufficiently many new nodes have been added in this way.

4. **Load Balancing Properties**

   Because the data items are searched for along a path from a random top supernode to the bottom supernodes containing the item, and because these bottom supernodes are chosen at random, the load will be well balanced as long as the number of requests for different data items is itself balanced. This follows because a uniform distribution on the search for data items translates to a uniform distribution on top to bottom paths through the butterfly.

4 Proofs

4.1 Proof Overview

Technically, the proof makes extensive use of random constructions and the Probabilistic Method [2].

We first show that with high probability, all but an arbitrarily small constant times $n/\log n$ of the supernodes are good, where good means that (a) they have $O(\log n)$ nodes associated with them, and, (b) they have $\Omega(\log n)$ live nodes after adversarial deletion. This implies that all but a small constant fraction of the paths through the butterfly contain only good supernodes.

Search is preformed by broadcasting the search to all the nodes in (a constant number of) top supernodes, followed by a sequence of broadcasts between every successive pair of supernodes along the paths between one of these top supernodes and a constant number of bottom supernodes. Fix one such path. The broadcast between two successive supernodes along the path makes use of the expander graph connecting these two supernodes. When we broadcast from the live nodes in a supernode to the following supernode, the nodes that we reach may be both live and dead(see Figure 3).

Assume that we broadcast along a path, all of whose supernodes are good. One problem is that we are not guaranteed to reach all the live nodes in the next supernode along the path. Instead, we reduce our requirements to ensure that at every stage, we reach at least $\delta \log n$ live nodes, for some constant δ. The crucial observation is that if we broadcast from $\delta \log n$ live nodes in one supernode, we are guaranteed

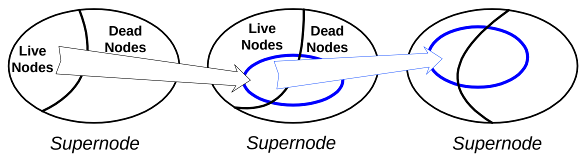

Figure 3: Traversal of a path through the butterfly.

to reach at least $\delta \log n$ live nodes in the subsequent supernode, with high probability. This follows by using the expansion properties of the bipartite expander connection between two successive supernodes.

Recall that the nodes are connected to a constant number of random top supernodes, and that the data items are stored in a constant number of random bottom supernodes. The fact that we can broadcast along all but an arbitrarily small fraction of the paths in the butterfly implies that most of the nodes can reach most of the content.

In several statements of the lemmata and theorems in this section, we require that n, the number of nodes in the network, be sufficiently large to get our result. We note that, technically, this requirement is not necessary since if it fails then n is a constant and our claims trivially hold.

4.2 Definitions

Definition 4.1. *A top or middle supernode is said to be (α, β)-good if it has at most $\beta \log n$ nodes mapped to it and at least $\alpha \log n$ nodes which are not under control of the adversary.*

Definition 4.2. *A bottom supernode is said to be (α, β)-good if it has at most $\beta \log n$ nodes mapped to it and at least $\alpha \log n$ nodes which are not under control of the adversary and if there are no more than $\beta B \ln n$ data items that map to the node.*

Definition 4.3. *An (α, β)-good path is a path through the butterfly network from a top supernode to a bottom supernode all of whose supernodes are (α, β)-good supernodes.*

Definition 4.4. *A top supernode is called (γ, α, β)-expansive if there exist $\gamma n / \log n$ (α, β)-good paths that start at this supernode.*

4.3 Technical Lemmata Following are three technical lemmata about bipartite expanders that we will use in our proofs. The proof of the first lemma is well known [21] (see also [17]) and the proof of the next two lemmata are slight variants on the proof of the first. Due to space constraints, the proofs of all three lemmata are omitted from this extended abstract.

Lemma 4.1. *Let l, r, l', r', d and n be any positive values where $l' \leq l$ and $r' \leq r$ and*

$$d \geq \frac{r}{r'l'} \left(l' \ln \left(\frac{le}{l'} \right) + r' \ln \left(\frac{re}{r'} \right) + 2 \ln n \right).$$

Let G be a random bipartite multigraph with left side L and right side R where $|L| = l$ and $|R| = r$ and each node in L has edges to d random neighbors in R. Then with probability at least $1 - 1/n^2$, any subset of L of size l' shares an edge with any subset of R of size r' .

Lemma 4.2. *Let l, r, l', r', d, λ and n be any positive values where $l' \leq l$, $r' \leq r$, $0 < \lambda < 1$ and*

$$d \geq \frac{2r}{r'l'(1-\lambda)^2} \left(l' \ln \left(\frac{le}{l'} \right) + r' \ln \left(\frac{re}{r'} \right) + 2 \ln n \right).$$

Let G be a random bipartite multigraph with left side L and right side R where $|L| = l$ and $|R| = r$ and each node in L has edges to d random neighbors in R. Then with probability at least $1 - 1/n^2$, for any set $L' \subset L$ where $|L'| = l'$, there is no set $R' \subset R$, where $|R'| = r'$ such that all nodes in R' share less than $\lambda l' d / r$ edges with L'.

Lemma 4.3. *Let l, r, r', d, β' and n be any positive values where $l' \leq l$, $\beta' > 1$ and*

$$d \geq \frac{4r}{r'l(\beta'-1)^2} \left(r' \ln \left(\frac{re}{r'} \right) + 2 \ln n \right).$$

Let G be a random bipartite multigraph with left side L and right side R where $|L| = l$ and $|R| = r$ and each node in L has edges to d random neighbors in R. Then with probability at least $1 - 1/n^2$, there is no set $R' \subset R$, where $|R'| = r'$ such that all nodes in R' have degree greater than $\beta' l d / r$.

4.4 (α, β)-good Supernodes

Lemma 4.4. *Let α, δ', n be values where $\alpha < 1/2$ and $\delta' > 0$ and let $k(\delta', \alpha)$ be a value that depends only on α, δ' and assume n is sufficiently large. Let each node participate in $k(\delta', \alpha) \ln n$ random middle supernodes. Then removing any set of $n/2$ nodes still leaves all but $\delta' n / \ln n$ middle supernodes with at least $\alpha k(\delta', \alpha) \ln n$ live nodes.*

Proof. For simplicity, we will assume there are n middle supernodes (we can throw out any excess supernodes).

Let $l = n$, $l' = n/2$, $r = n$, $r' = \delta' n / \ln n$, $\lambda = 2\alpha$ and $d = k(\delta', \alpha) \ln n$ in Lemma 4.2. We want probability less than $1/n^2$ of being able to remove $n/2$ nodes and having a set of $\delta' n / \ln n$ supernodes all with less than $\alpha k(\delta', \alpha) \ln n$ live nodes. This happens provided that the number of connections from each supernode is bounded as in Lemma 4.2 which happens when:

$$k(\delta', \alpha) \geq \frac{2 \ln(2e)}{\delta'(1-2\alpha)^2} + o(1).$$

Lemma 4.5. *Let β, δ', n, k be values such that $\beta > 1$, $\delta' > 0$ and assume n is sufficiently large. Let each node participate in $k \ln n$ of the middle supernodes, chosen uniformly at random. Then all but $\delta' n / \ln n$ middle supernodes have less than $\beta k \ln n$ participating nodes with probability at least $1 - 1/n^2$.*

Proof. For simplicity, we will assume there are n middle supernodes (we can throw out any excess supernodes and the lemma will still hold). Let $l = n$, $r = n$, $r' = \delta' n / \ln n$, $d = k \ln n$ and $\beta' = \beta$ in Lemma 4.3. Then the statement in this lemma holds provided that:

$$k \geq \frac{4}{(\beta - 1)^2 \ln n} \cdot \ln \left(\frac{\ln n}{\delta'} + \frac{2}{\delta' n} \right).$$

The right hand side of this equation goes to 0 as n goes to infinity.

Lemma 4.6. *Let α, δ', n be values such that $\alpha < 1/2$, $\delta' > 0$ and let $k(\delta', \alpha)$ be a value that depends only on δ' and α and assume n is sufficiently large. Let each node participate in $k(\delta', \alpha)$ top (bottom) supernodes. Then removing any set of $n/2$ nodes still leaves all but $\delta' n / \ln n$ top (bottom) supernodes with at least $\alpha k(\delta', \alpha) \ln n$ live nodes.*

Proof. Let $l = n$, $l' = n/2$, $r = n / \ln n$, $r' = \delta' n / \ln n$, $\lambda = 2\alpha$ and $d = k(\delta', \alpha)$ in Lemma 4.2. We want probability less than $1/n^2$ of being able to remove $n/2$ nodes and having a set of $\delta' n / \ln n$ supernodes all with less than $\alpha k(\delta', \alpha) \ln n$ live nodes. We get this provided that the number of connections from each supernode is bounded as in Lemma 4.2:

$$k(\delta', \alpha) = \frac{2 \ln(2e)}{\delta' (1 - 2\alpha)^2} + o(1).$$

Lemma 4.7. *Let β, δ', n, k be values such that $\beta > 1$, $\delta' > 0$ and n is sufficiently large. Let each node participate in k of the top (bottom) supernodes (chosen uniformly at random). Then all but $\delta' n / \ln n$ top (bottom) supernodes consist of less than $\beta k \ln n$ nodes with probability at least $1 - 1/n^2$.*

Proof. Let $l = n$, $r = n / \ln n$, $r' = \delta' n / \ln n$, $d = k$ and $\beta' = \beta$ in Lemma 4.3. Then the statement in this lemma holds provided that:

$$k \geq \frac{4}{\ln n (\beta - 1)^2} \cdot \left(\ln \left(\frac{e}{\delta'} \right) + \frac{2 \ln n}{\delta' n} \right).$$

The right hand side of this equation goes to 0 as n goes to infinity.

Corollary 4.1. *Let β, δ', n, k be values such that $\beta > 1$, $\delta' > 0$ and n is sufficiently large. Let each data item be stored in k of the bottom supernodes (chosen uniformly at random). Then all but $\delta' n / \ln n$ bottom supernodes have less than $\beta k \ln n$ data items stored on them with probability at least $1 - 1/n^2$.*

Proof. Let the data items be the left side of a bipartite graph and the bottom supernodes be the right side. The proof is then the same as Lemma 4.7.

Corollary 4.2. *Let $\delta' > 0$, $\alpha < 1/2$, $\beta > 1$. Let $k(\delta', \alpha)$, be a value depending only on δ' and assume n is sufficiently large. Let each node appear in $k(\delta', \alpha)$ top supernodes, $k(\delta', \alpha)$ bottom supernodes and $k(\delta', \alpha) \ln n$ middle supernodes. Then all but $\delta' n$ of the supernodes are $(\alpha k(\delta', \alpha), \beta k(\delta', \alpha))$-good with probability $1 - O(1/n^2)$.*

Proof. Use

$$k(\delta', \alpha) = \frac{10}{3} \cdot \frac{2 \ln(2e)}{\delta' (1 - 2\alpha)^2}$$

in Lemma 4.6. Then we know that no more than $3\delta' n / (10 \ln n)$ top supernodes and no more than $3\delta' n / (10 \ln n)$ bottom supernodes have less than $\alpha k(\delta', \alpha) \ln n$ live nodes. Next plugging $k(\delta', \alpha)$ into Lemma 4.4 gives that no more than $3\delta' n / (10 \ln n)$ middle supernodes have less than $\alpha k(\delta', \alpha) \ln n$ live nodes.

Next using $k(\delta', \alpha)$ in Lemma 4.7 and Lemma 4.5 gives that no more than $\delta' n / (20 \ln n)$ of the supernodes can have more than $\beta k(\delta', \alpha) \ln n$ nodes in them. Finally, using $k(\delta', \alpha)$ in Lemma 4.1 gives that no more than $\delta' n / (20 \ln n)$ of the bottom supernodes can have more than $\beta k(\delta', \alpha) \ln n$ data items stored at them. If we put these results together, we get that no more than $\delta n / \ln n$ supernodes are not $(\alpha k(\delta', \alpha), \beta k(\delta', \alpha))$-good with probability $1 - O(1/n^2)$

4.5 (γ, α, β)-expansive Supernodes

Theorem 4.1. *Let $\delta > 0$, $\alpha < 1/2$, $0 < \gamma < 1$, $\beta > 1$. Let $k(\delta, \alpha, \gamma)$ be a value depending only on δ, α, γ and assume n is sufficiently large. Let each node participate in $k(\delta, \alpha, \gamma)$ top supernodes, $k(\delta, \alpha, \gamma)$ bottom supernodes and $k(\delta, \alpha, \gamma) \ln n$ middle supernodes. Then all but $\delta n / \ln n$ top supernodes are $(\gamma, \alpha k(\delta, \alpha), \beta k(\delta, \alpha))$-expansive with probability $1 - O(1/n^2)$.*

Proof. Assume that for some particular $k(\delta, \alpha, \gamma)$ that more than $\delta n / \ln n$ top supernodes are not $(\gamma, \alpha k(\delta, \alpha, \gamma), \beta k(\delta, \alpha, \gamma)$-expansive. Then each of these bad top supernodes has $(1 - \gamma n) / \ln n$ paths that are not $(\alpha k(\delta, \alpha, \gamma), \beta k(\delta, \alpha, \gamma))$-good. So the total number of paths that are not $(\alpha k(\delta, \alpha, \gamma), \beta k(\delta, \alpha, \gamma))$-good is more than

$$\frac{\delta (1 - \gamma) n^2}{\ln^2 n}.$$

We will show there is a $k(\delta, \alpha, \gamma)$ such that this event will not occur with high probability. Let $\delta' = \delta(1 - \gamma)$ and let

$$k(\delta, \alpha, \gamma) = \frac{10}{3} \cdot \frac{2 \ln(2e)}{\delta (1 - \gamma)(1 - 2\alpha)^2}.$$

Then we know by Lemma 4.2 that with high probability, there are no more than $\delta(1-\gamma)n/\ln n$ supernodes that are not $(\alpha k(\delta,\alpha,\gamma),\beta k(\delta,\alpha,\gamma))$good. We also know that each of these supernodes which are not good cause at most $n/\ln n$ paths in the butterfly to be not $(\alpha k(\delta,\alpha,\gamma),\beta k(\delta,\alpha,\gamma))$-good. Hence the number of paths that are not $(\alpha k(\delta,\alpha,\gamma),\beta k(\delta,\alpha,\gamma))$-good is no more than $\delta(1-\gamma)n^2/(\ln^2 n)$ which is what we wanted to show.

4.6 (α,β)-good Paths to Data Items We will use the following lemma to show that almost all the nodes are connected to some appropriately expansive top supernode.

Lemma 4.8. *Let $\delta > 0$, $\epsilon > 0$ and n be sufficiently large. Then exists a constant $k(\delta,\epsilon)$ depending only on ϵ and δ such that if each node connects to $k(\delta,\epsilon)$ random top supernodes then with high probability, any subset of the top supernodes of size $(1-\delta)n/\ln n$ can be reached by at least $(1-\epsilon)n$ nodes.*

Proof. We imagine the n nodes as the left side of a bipartite graph and the $n/\ln n$ top supernodes as the right side and an edge between a node and a top supernode in this graph if and only if the node and supernode are connected.

For the statement in the lemma to be false, there must be some set of ϵn nodes on the left side of the graph and some set of $(1-\delta)n/\ln n$ top supernodes on the right side of the graph that share no edge. We can find $k(\delta,\epsilon)$ large enough that this event occurs with probability no more than $1/n^2$ by plugging in $l = n$, $l' = \epsilon n$, $r = n/\ln n$ and $r' = (1-\delta)(n/\ln n)$ into Lemma 4.1. The bound found is:

$$k(\delta,\epsilon) \geq \frac{\ln\left(\frac{e}{\epsilon}\right)}{1-\delta} + o(1).$$

We will use the following lemma to show that if we can reach γ bottom supernodes that have some live nodes in them that we can reach most of the data items.

Lemma 4.9. *Let γ, n, ϵ be any positive values such that $\epsilon > 0$, $\gamma > 0$. There exists a $k(\epsilon,\gamma)$ which depends only on ϵ,γ such that if each bottom supernode holds $k(\epsilon,\gamma)\ln n$ random data items, then any subset of bottom supernodes of size $\gamma n/\ln n$ holds $(1-\epsilon)n$ unique data items.*

Proof. We imagine the n data items as the left side of a bipartite graph and the $n/\ln n$ bottom supernodes as the right side and an edge between a data item and a bottom supernode in this graph if and only if the supernode contains the data item. The bad event is that there is some set of $\gamma n/\ln n$ supernodes on the right that share no edge with some set of ϵn data items on the right. We can find $k(\epsilon,\gamma)$ large enough that this event occurs with probability no more than $1/n^2$ by plugging in $l = n$, $l' = \epsilon n$ into $r = n/\ln n$, $r' = \gamma n/\ln n$ into Lemma 4.1. We get

$$k(\epsilon,\gamma) \geq \frac{1}{\gamma}\cdot\ln\frac{e}{\epsilon} + o(1)$$

4.7 Connections between (α,β)-good supernodes

Lemma 4.10. *Let α,β,α',n be any positive values where $\alpha' < \alpha$, $\alpha > 0$ and let C be the number of supernodes to which each node connects. Let X and Y be two supernodes that are both $(\alpha C,\beta C)$-good. Let each node in X have edges to $k(\alpha,\beta,\alpha')$ random nodes in Y where $k(\alpha,\beta,\alpha')$ is a value depending only on α,β and α'. Then with probability at least $1-1/n^2$, any set of $\alpha' C\ln n$ nodes in X has at least $\alpha' C\ln n$ live neighbors in Y*

Proof. Consider the event where there is some set of $\alpha' C\ln n$ nodes in X which do not have $\alpha' C\ln n$ live neighbors in Y. There are $\alpha C\ln n$ live nodes in Y so for this event to happen, there must be some set of $(\alpha-\alpha')C\ln n$ live nodes in Y that share no edge with some set of $\alpha' d\ln n$ nodes in X. We note that the probability that there are two such sets which share no edge is largest when X and Y have the most possible nodes. Hence we will find a $k(\alpha,\beta,\alpha')$ large enough to make this bad event occur with probability less than $1/n^2$ if in Lemma 4.1 we set $l = \beta C\ln n$, $r = \beta C\ln n$, $l' = \alpha' C\ln n$ and $r' = (\alpha-\alpha')C\ln n$. When we do this, we get that $k(\alpha,\beta,\alpha')$ must be greater than or equal to:

$$\left(\frac{\beta}{\alpha'(\alpha-\alpha')}\right)\cdot\left(\alpha'\ln\left(\frac{\beta e}{\alpha'}\right) + (\alpha-\alpha')\ln\left(\frac{\beta e}{\alpha-\alpha'}\right) + \frac{2}{C}\right).$$

4.8 Putting it All Together We are now ready to give the proof of Theorem 3.1.

Proof. Let $\delta,\alpha,\gamma,\alpha',\beta$ be any values such that $0 < \delta < 1$, $0 < \alpha < 1/2$, $0 < \alpha' < \alpha$, $\beta > 1$ and $0 < \gamma < 1$. Let

$$C = \tfrac{10}{3}\cdot\tfrac{2\ln(2e)}{\delta(1-\gamma)(1-2\alpha)^2};$$
$$T = \tfrac{\ln\left(\frac{e}{\epsilon}\right)}{1-\delta};$$
$$B = \tfrac{1}{\gamma}\ln\left(\tfrac{e}{\epsilon}\right);$$
$$D = \left(\tfrac{\beta}{\alpha'(\alpha-\alpha')}\right)\left(\alpha'\ln\left(\tfrac{\beta e}{\alpha'}\right) + (\alpha-\alpha')\ln\left(\tfrac{\beta e}{\alpha-\alpha'}\right) + \tfrac{2}{C}\right)$$

Let each node connect to C top, C bottom and C middle supernodes. Then by Theorem 4.1, at least $(1-\delta)n/\ln n$ top supernodes are $(\gamma,\alpha C,\beta C)$-expansive. Let each node connect to T top supernodes. Then by Lemma 4.8, at least $(1-\epsilon)n$ nodes can connect to some $(\gamma,\alpha C,\beta C)$-expansive top supernode. Let each data item map to B bottom supernodes. Then by Lemma 4.9, at least $(1-\epsilon)n$ nodes have $(\alpha C, \beta C)$-good paths to at least $(1-\epsilon)n$ data items.

Finally, let each node in a middle supernode have D random connections to nodes in neighboring supernodes in the butterfly network. Then by Lemma 4.10, at least $(1-\epsilon)n$ nodes can broadcast to enough bottom supernodes so that they can reach at least $(1-\epsilon)n$ data items.

Each node requires T links to connect to the top supernodes; $2D$ links for each of the C top supernodes it plays a role in; $2D$ links for each of the $C \ln n$ middle supernodes it plays a role in and $B\beta \ln n$ storage for each of the C bottom supernodes it plays a role in. The total amount of memory required is thus

$$T + 2DC + C \ln n(2D + B\beta),$$

which is less than $k_1(\epsilon) \log n$ for some $k_1(\epsilon)$ dependent only on ϵ.

Our search algorithm will find paths to at most B bottom supernodes for a given data item and each of these paths has less than $\log n$ hops in it so the search time is no more than

$$k_2(\epsilon) \log n = B \log n.$$

Each supernode contains no more that $\beta \ln n$ nodes and in each search, exactly T top supernodes send no more than B messages so the total number of messages transmitted during a search is no more than

$$k_3(\epsilon) \log^2 n = (TB\beta C) \log^2 n.$$

5 Modifications for Spam Resistant Content Addressable Network

We only sketch the changes in the network and the proofs to allow a spam resistant content addressable network. The arguments are based on slight modifications to the proofs of section 4.

The first modification is that rather than have a constant degree expander between two supernodes connected in the butterfly, we will have a full bipartite graph between the nodes of these two supernodes. Since we've insisted that the total number of adversary controlled nodes be strictly less than $n/2$, we can guarantee that a $1-\epsilon$ fraction of the paths in the butterfly have all supernodes with a majority of good (non-adversary controlled) nodes. In particular, by substituting appropriate values in Lemma 4.2 and Lemma 4.3 we can guarantee that all but $\epsilon n / \log n$ of the supernodes have a majority of good nodes. This then implies that no more than an ϵ fraction of the paths pass through such "adversary-majority" supernodes. As before, this implies that most of the nodes can access most of the content through paths that don't contain any "adversary-majority" supernodes.

Search is performed as with the original construction, and after the bottommost supernodes are reached, the data content flows back along the same links as the search went down. We modify the protocol so that along this return flow, every node passes up a data value only if a majority of the values it received from the nodes below it are the same. This means that if there are no "adversary-majority" supernodes on the path, then all good nodes will take a majority value from a set in which good nodes are a majority. Thus, along such a path, only the correct data value will be passed upwards by good nodes. At the top, the node that issued the search takes the majority value amongst the $(O(\log n))$ values it receives as the final search result.

To summarize, the main theorem for spam resistant content addressable networks is as follows:

THEOREM 5.1. *For any constant $c < 1/2$ such that the adversary controls no more than cn nodes, and for all $\epsilon > 0$, there exist constants $k_1(\epsilon)$, $k_2(\epsilon)$, $k_3(\epsilon)$ which depend only on ϵ such that*

- *Every node requires $k_1(\epsilon) \log^2 n$ memory.*
- *Search for a data item takes no more than $k_3(\epsilon) \log n$ time. (This is under the assumption that network latency overwhelms processing time for one message, otherwise the time is $O(\log^2 n)$.)*
- *Search for a data item requires no more than $k_3(\epsilon) \log^3 n$ messages.*
- *All but ϵn nodes can search successfully for all but ϵn of the true data items.*

6 Discussion and Open Problems

We conclude with some open issues:

1. For the deletion resistant content addressable network: Is there a mechanism for dynamically maintaining our network when large numbers of nodes are deleted or added to the network? Is it possible to reduce the number of messages that are sent in a search for a data item from $O(\log^2 n)$ to $O(\log n)$?
2. Can one improve on the construction for the spam resistant content addressable network?
3. Can one deal efficiently with more general Byzantine faults? For example, the adversary could use nodes under his control to flood the network with irrelevant searches, this is not dealt with by either of our solutions.

4. We conjecture that our network has the property that it is poly-log competitive with any fixed degree network. *I.e.*, we conjecture that given any fixed degree network topology, where n items are distributed amongst n nodes, and any set of access requests that can be dealt with fixed sized buffers, then our network will also deal with the same set of requests by introducing no more than a polylog slowdown.

7 Acknowledgments

We gratefully thank Anna Karlin, Prabhakar Raghavan, Stefan Saroiu, and Steven Gribble for their great help with this paper.

References

[1] Noga Alon, Haim Kaplan, Michael Krivelevich, Dahlia Malkhi, and Julien Stern. Scalable secure storage when half the system is faulty. In *Proceedings of the 27th International Colloquium on Automata, Languages and Programming*, 2000.

[2] Noga Alon and Joel Spencer. *The Probabilistic Method, 2nd Edition.* John Wiley & Sons, 2000.

[3] Yonatan Aumann and Michael Bender. Fault tolerant data structures. In *IEEE Symposium on Foundations of Computer Science*, 1996.

[4] M. Bellare and P. Rogaway. Random oracles are practical: a paradigm for designing efficient protocols. In *The First ACM Conference on Computer and Communications Security*, pages 62–73, 1993.

[5] John Borland. Gnutella girds against spam attacks. *CNET News.com*, August 2000. http://news.cnet.com/news/0-1005-200-2489605.html.

[6] Clip2. Gnutella: To the bandwidth barrier and beyond. http://dss.clip2.com/gnutella.html.

[7] Electronic Freedom Foundation. Eff — censorship — internet censorship legislation & regulation (cda, etc.) — archive. http://www.eff.org/-pub/Censorship/Internet_censorship_bills.

[8] D.K. Gifford. Weighted voting for replicated data. In *Proc. of the Seventh ACM Symposium on Operating Systems Principles*, pages 150–159, 1979.

[9] J. Hastad, T. Leighton, and M. Newman. Fast computation using faulty hypercubes. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, 1989.

[10] Anna R. Karlin, Greg Nelson, and Hisao Tamaki. On the fault tolerance of the butterfly. In *ACM Symposium on Theory of Computing*, 1994.

[11] John Kubiatowicz, David Bindel, Yan Chen, Steven Czerwinski, Patrick Eaton, Dennis Geels, Ramakrishna Gummadi, Sean Rhea, Hakim Weatherspoon, Westley Weimer, Chris Wells, and Ben Zhao. Oceanstore: An architecture for global-scale persistent storage. In *Appears in Proceedings of the Ninth international Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000)*, 2000.

[12] Thomson Leighton, Bruce Maggs, and Ramesh Sitamaran. On the fault tolerance of some popular bounded-degree networks. *SIAM Journal on Computing*, 1998.

[13] Dahlia Malkhi, Michael Reiter, and Avishai Wool. The load and availability of byzantine quorum systems. *SIAM Journal of Computing*, 29(6):1889–1906, 2000.

[14] Dahlia Malkhi, Michael Reiter, Avishai Wool, and Rebecca N. Wright. Probabilistic byzantine quorum systems. In *Symposium on Principles of Distributed Computing*, 1998.

[15] Aviel D. Rubin Marc Waldman and Lorrie Faith Cranor. Publius: A robust, tamper-evident, censorship-resistant, web publishing system. In *Proc. 9th USENIX Security Symposium*, pages 59–72, August 2000.

[16] Robert Marquand. China's web users kept on their toes. http://www.csmonitor.com/durable/-2000/12/07/fp7s1-csm.shtml.

[17] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[18] Index on Censorship. Index on censorship homepage. http://www.indexoncensorship.org.

[19] Andy Oram, editor. *Peer-to-Peer: Harnessing the Power of Disruptive Technologies.* O'Reilly & Associates, July 2001.

[20] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. Building low-diameter p2p networks. In *STOC 2001, Crete, Greece*, 2001.

[21] M. Pinsker. On the complexity of a concentrator. In *7th International Teletraffic Conference*, 1973.

[22] C. Plaxton, R. Rajaram, and A.W. Richa. Accessing nearby copies of replicated objects in a distributed environment. In *Proceedings of the Ninth Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, 1997.

[23] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A Scalable Content-Addressable Network. In *Proceedings of the ACM SIGCOMM 2001 Technical Conference*, San Diego, CA, USA, August 2001.

[24] Stefan Saroiu, P. Krishna Gummadi, and Steven D. Gribble. A Measurement Study of Peer-to-Peer File Sharing Systems. In *Proceedings of Multimedia Computing and Networking*, 2002.

[25] Adi Shamir. How to share a secret. *Communications of the ACM, 22,pp. 612–613*, 1979.

[26] Ion Stoica, Robert Morris, David Karger, Frans Kaashoek, and Hari Balakrishnan. Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications. In *Proceedings of the ACM SIGCOMM 2001 Technical Conference*, San Diego, CA, USA, August 2001.

[27] Gnutella website. http://gnutella.wego.com/.

[28] Napster website. http://www.napster.com/.

[29] B.Y. Zhao, K.D. Kubiatowicz, and A.D. Joseph. Tapestry: An Infrastructure for Fault-Resilient Wide-Area Location and Routing. Technical Report UCB//CSD-01-1141, University of California at Berkeley Technical Report, April 2001.