Search and Replication in Unstructured Peer-to-Peer Networks

Qin Lv
Dept. of Computer Science
Princeton University
qlv@CS.Princeton.EDU

Pei Cao
Cisco Systems, Inc.
cao@cisco.com

Edith Cohen
AT&T Labs-Research
edith@research.att.com

Kai Li
Dept. of Computer Science
Princeton University
li@CS.Princeton.EDU

Scott Shenker
ICSI
shenker@icsi.berkeley.edu

ABSTRACT

Decentralized and unstructured peer-to-peer networks such as Gnutella are attractive for certain applications because they require no centralized directories and no precise control over network topology or data placement. However, the flooding-based query algorithm used in Gnutella does not scale; each query generates a large amount of traffic and large systems quickly become overwhelmed by the query-induced load. This paper explores, through simulation, various alternatives to Gnutella's query algorithm, data replication strategy, and network topology. We propose a query algorithm based on multiple random walks that resolves queries almost as quickly as Gnutella's flooding method while reducing the network traffic by two orders of magnitude in many cases. We also present simulation results on a distributed replication strategy proposed in [8]. Finally, we find that among the various network topologies we consider, uniform random graphs yield the best performance.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: [Online Information Services, Data Sharing]

General Terms

Algorithms, Measurement, Performance, Theory

Keywords

peer-to-peer, search, replication, unstructured

1. INTRODUCTION

The computer science community has become accustomed to the Internet's continuing rapid growth, but even to such jaded observers the explosive increase in Peer-to-Peer (P2P) network usage has been astounding. Within a few months of Napster's [16] introduction in 1999 the system had spread widely, and recent measurement data suggests that P2P applications are having a very significant and rapidly growing impact on Internet traffic [12, 17]. It is important to study the performance and scalability of these P2P networks.

Currently, there are several different architectures for P2P networks:

Centralized: Napster and other similar systems have a constantly-updated directory hosted at central locations (*e.g.*, the Napster web site). Nodes in the P2P network issue queries to the central directory server to find which other nodes hold the desired files. Such centralized approaches do not scale well and have single points of failure.

Decentralized but Structured: These systems have no central directory server, but they have a significant amount of structure. By "structure" we mean that the P2P overlay topology (that is, the set of connections between P2P members) is tightly controlled and that files are placed not at random nodes but at specified locations that will make subsequent queries easier to satisfy. In "loosely structured" systems this placement of files is based on hints; the Freenet P2P network [11] is an example of such systems. In "highly structured" systems both the P2P network topology and the placement of files are precisely determined; this tightly controlled structure enables the system to satisfy queries very efficiently. There is a growing literature on highly structured P2P systems which support a hash-table-like interface; see [18, 23, 20, 26]. Such highly structured P2P designs are quite prevalent in the research literature, but almost completely invisible on the current network. Moreover, it isn't clear how well such designs work with an extremely transient population of nodes, which seems to be a characteristic of the Napster community.

Decentralized and Unstructured: These are systems in which there is neither a centralized directory nor any precise control over the network topology or file placement. Gnutella [13] is an example of such designs. The

ICS'02, June 22-26, 2002, New York, New York, USA

network is formed by nodes joining the network following some loose rules (for example, those described in [6]). The resultant topology has certain properties, but the placement of files is not based on any knowledge of the topology (as it is in structured designs). To find a file, a node queries its neighbors. The most typical query method is flooding, where the query is propagated to all neighbors within a certain radius [6]. These unstructured designs are extremely resilient to nodes entering and leaving the system. However, the current search mechanisms are extremely unscalable, generating large loads on the network participants.

In this paper, we focus on Gnutella-like decentralized, unstructured P2P systems. We do so because (1) these systems are actively used by a large community of Internet users [24, 7], and (2) these systems have not yet been subject to much serious research, except for empirical studies [21, 22].

The goal of this paper is to study more-scalable alternatives to existing Gnutella algorithms, focusing on the search and replication aspects. We first quantify the poor scaling properties of the flooding search algorithms. We then propose, as an alternative, a k-walker random walk algorithm that greatly reduces the load generated by each query. We also show that active replication (where the files may be stored at arbitrary nodes) produces lower overall query load than non-active node-based replication (*i.e.*, a file is only replicated at nodes requesting the file). Confirming the results in [8], we use simulations to show that path replication, where the file is replicated along the path from the requester to the provider, yields a close-to-optimal replication distribution. Finally, we show that for unstructured networks, power-law random graphs are less desirable than uniform random graphs and so P2P systems should adopt graph-building algorithms that reduce the likelihood of very-high degree nodes.

2. METHODOLOGY

It is impossible to model all the dynamics of an Internet-based P2P system. In this paper, we are not trying to resolve small quantitative disparities between different algorithms, but instead are trying to reveal fundamental qualitative differences. While our simple models do not capture all aspects of reality, we hope they capture the essential features needed to understand the qualitative differences.

2.1 Abstractions

We look at three aspects of a P2P system: P2P network topology, query distribution and replication. By network topology, we mean the graph formed by the P2P overlay network; each P2P member has a certain number of "neighbors" and the set of neighbor connections forms the P2P overlay network. In this paper when we refer to the "network" we are referring to the P2P network, not the underlying Internet. For simplicity, we assume that the P2P network graph does not change during the simulation of our algorithms. By query distribution, we mean the distribution of query frequencies for individual files. Again, we assume that this distribution is fixed during our simulations. By replication, we mean the number of nodes that have a particular file.[1]

	#nodes	total #links	avg. node degree	std. dev.	max degree	median degree
PLRG	9230	20599	4.46	27.9	1746	1
Random	9836	20099	4.09	1.95	13	4
Gnutella	4736	13022	5.50	10.7	136	2
Grid	10000	19800	3.96	0.20	4	4

Table 1: Key statistics of the network topologies

Our assumption of fixed network topology and fixed query distributions are obviously gross simplifications. However, for the purpose of our study, if one assumes that the time to complete a search is short compared to the time of change in network topology and in query distribution, results obtained from these settings are still indicative of performance in real systems.

We use four network topologies in our study:

- Power-Law Random Graph (PLRG): this is a 9230-node random graph. The node degrees follow a power-law distribution: when ranked from the most connected to the least connected, the i'th most connected node has ω/i^α neighbors, where ω is a constant. Once the node degrees are chosen, the nodes are connected randomly; see [3]. Many real-life P2P networks have topologies that are power-law random graphs [14].

- Normal Random Graph (Random): a 9836-node random graph generated by a modified version of GT-ITM topology generator [5].

- Gnutella graph (Gnutella): the Gnutella network topology, as obtained in October 2000. The graph has 4736 nodes.[2] Its node degrees roughly follow a two-segment power-law distribution, as shown in Figure 1.

- Two-Dimensional Grid (Grid): a two-dimension (100x 100) grid. We choose this simple graph for comparison purposes.

Node degree information of the four graphs are shown in Figure 1. and are also summarized in Table 1.

We assume that there are m objects of interest. (In this paper we use the terms "file" and "object" interchangeably.) Let q_i be the relative popularity, in terms of the number of queries issued for it, of the i'th object. The values are normalized:

$$\sum_{i=1}^{m} q_i = 1 \tag{1}$$

We investigate the following distributions:

- Uniform: all objects are equally popular.

$$\text{Uniform} : q_i = 1/m \tag{2}$$

- Zipf-like: object popularity follows a Zipf-like distribution. Studies have shown that Napster, Gnutella and Web queries follow Zipf-like distributions [4, 21].

$$\text{Zipf-like} : q_i \propto 1/i^\alpha \tag{3}$$

[1] We will use replication ratio to mean the *percentage* of nodes having the file.

[2] Cautions are taken when we choose metrics and parameters in our analysis and simulations to account for the difference in network sizes.

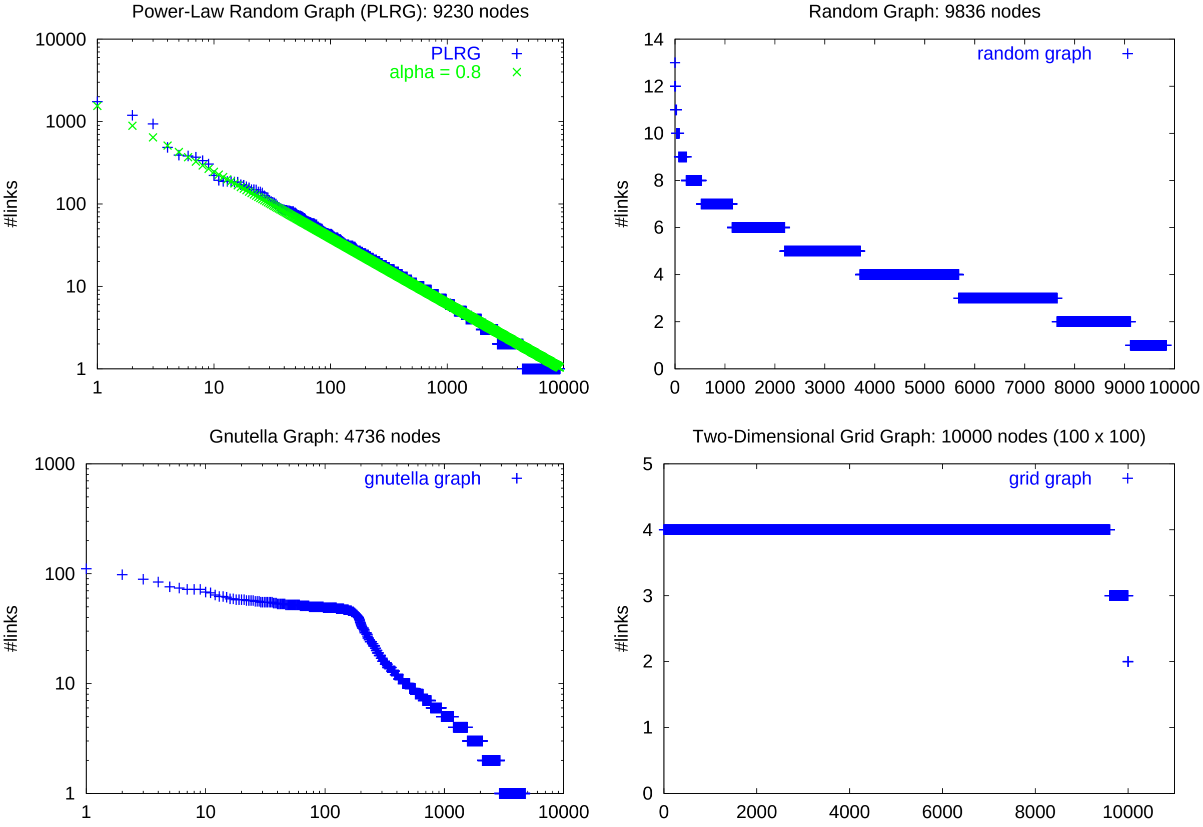


Figure 1: Distribution of node degrees in the four network topology graphs. Note that we use log scale for PLRG and Gnutella, and linear scale for Random and Grid.

We assume each object i is replicated on r_i nodes, and the total number of objects stored in the network is R.

$$\sum_{i=1}^{m} r_i = R \tag{4}$$

We consider several static replication distributions: [3]

- Uniform: all objects are replicated at the same number of nodes.

$$\text{Uniform}: r_i = R/m \tag{5}$$

- Proportional: the replication of an object i is proportional to the query probability of the object. If only nodes requesting an object store the object, then the replication distribution is usually proportional to query distribution.

$$\text{Proportional}: r_i \propto q_i \tag{6}$$

- Square-root: replication of an object i is proportional to the square root of its query probability q_i. The reason for this distribution is discussed in [8] and reviewed in Section 5.

$$\text{Square-root}: r_i \propto \sqrt{q_i} \tag{7}$$

A simulation, in the static case, is determined by the combination of query distribution and replication distribution. When the replication is uniform, the query distribution is irrelevant (since all objects are replicated by the same amount, the search times are equivalent). When the query distribution is uniform, all three replication distributions are equivalent. Thus, there are only three relevant combinations of query distribution and replication distribution: uniform/uniform, Zipf-like/proportional, and Zipf-like/square-root.

2.2 Metrics

Performance issues in real P2P systems are extremely complicated. In addition to issues such as load on the network, load on network participants, and delays in getting positive answers, there are a host of other criteria such as success rate of the search, the bandwidth of the selected provider nodes, and fairness to both the requester and the provider. It is impossible for us to use all of these criteria in evaluating search and replication algorithms.

Instead, we focus on efficiency aspects solely, and use the following simple metrics in our abstract P2P networks. These metrics, though simple, reflect the fundamental properties of the algorithms.

- User aspects:
  - Pr(success): the probability of finding the queried object before the search terminates. Different al-

[3] We assume static replication when we study various search methods. We'll look at dynamic replication when we investigate different replication strategies.

gorithms have different criteria for terminating a search, leading to different probabilities of success under various replication distributions.

  - #hops: delay in finding an object as measured in number of hops. We do not model the actual network latency here, but rather just measure the abstract number of hops that a successful search message travels before it replies to the originator.

- Load aspects:
  - #msgs per node: overhead of an algorithm as measured in average number of search messages each node in the P2P network has to process. The motivation for this metric is that in P2P systems, the most notable overhead tends to be the processing load that the network imposes on each participant. The load, usually interrupt processing or message processing, is directly proportional to the number of messages that the node has to process.
  - #nodes visited: the number of P2P network participants that a query's search messages travel through. This is an indirect measure of the impact that a query generates on the whole network.
  - percentage of message duplication: calculated as (total_#msgs - #nodes visited)/total_#msgs.
  - peak #msgs: to identify hot spots in the network, we calculate the number of messages that the busiest node has to process for a set of queries.

- Aggregate performance: for each of the above measures, which are per-query measures, we calculate an aggregate performance number, which is the performance of each query convoluted with the query probability. That is, if the performance measure for object i is $a(i)$, then the aggregate is $\sum_i q_i * a(i)$.

2.3 Simulation Methodology

For each set of simulations, we first select the topology and the query/replication distributions. For each object i with replication r_i, we generate *numPlace* different sets of random replica placements (each set contains r_i random nodes, on which to place the replicas of object i). For each replica placement, we randomly choose *numQuery* different nodes from which to initiate a query for object i. Then, for each query, we simulate the searching process using the designated search method. We can run the simulation for each query independent of other queries because the object replication is fixed, and hence running all the queries concurrently is the same as running each one separately and then summing the results.

Statistics are collected from the *numPlace* × *numQuery* queries. For the results we present in this paper, *numPlace = 10* and *numQuery = 100*, resulting in 1000 different queries for each object. This ensures small standard deviation in our results.[4] We then calculate the aggregate results for the above metrics.

As a final note about our abstractions and metrics, we stress that they omit a lot of issues, including the true dynamics of node coming and going in the network, the message delays in the network, the actual load on a network node for processing and propagating messages, etc. However, these models help us understand the fundamental properties of various search and replication algorithms.

3. LIMITATIONS OF FLOODING

One of the major load issues in P2P networks is the load on individual network participants. Typically, the participants are PCs at home or office, and are used for normal work and entertainment. If a PC has to handle many network interrupts when it joins the P2P network, the user will be forced to take the PC off the P2P network to get "real" work done. This in turn limits the size and the usefulness of the P2P network. Unfortunately, the flooding search algorithm used in Gnutella exacerbates this problem.

Gnutella uses TTL(Time-To-Live) to control the number of hops that a query can be propagated. However, choosing the appropriate TTL is not easy. If the TTL is too high, the node unnecessarily burdens the network. If the TTL is too low, the node might not find the object even though a copy exists somewhere.

Figure 2 shows the probability of success and average per-node message overhead of flooding as TTL increases. The search is for an object that is replicated at 0.125% of the nodes. We can see from the figures that different TTLs are needed in different network topologies. And the values differ for different replication ratios. Unfortunately, since in practice the replication ratio of an object is unknown, users have to set TTLs high to ensure success of the query.

Another problem with flooding is that, there are many duplicate messages introduced by flooding, particularly in high connectivity graphs. By duplicate messages we mean the multiple copies of a query that are sent to a node by its multiple neighbors. Duplicate queries are pure overhead; they incur extra network interrupt processing at the nodes receiving them but do not contribute to increased chance of finding the object. Duplication detection mechanisms are always needed in flood-style search so that duplicate messages are detected and not forwarded. However, even with this duplicate suppression the number of duplicate messages in flooding-style algorithms can be excessive, and the problem worsens as the TTL increases.

Figure 3 shows the percentage of duplicate messages and the number of unique nodes visited as TTL increases. As we can see, when TTL increases, the number of unique nodes visited increases, but the percentage of duplicate messages also increases.

These limitations mean that flooding incurs considerable message processing overhead for each query, increasing the load on each node as the network expands and the query rate increases, to the point that a node can be so loaded that it has to leave the network. Other researchers have also noted the limitations of flooding [19].

Our simulations also show that Power-Law random graphs and Gnutella style graphs are particularly bad with flooding. The presence of highly connected nodes mean that its duplication ratios are much higher than those in the random graph, because many nodes' neighbors overlap. In fact, for flooding, the random graph would be the best topology because in a truly random graph the duplication ratio (likeli-

[4] Due to space limitation we do not present standard deviations associated with the averages, except to point out that the standard deviation data do not change the conclusions in the paper.

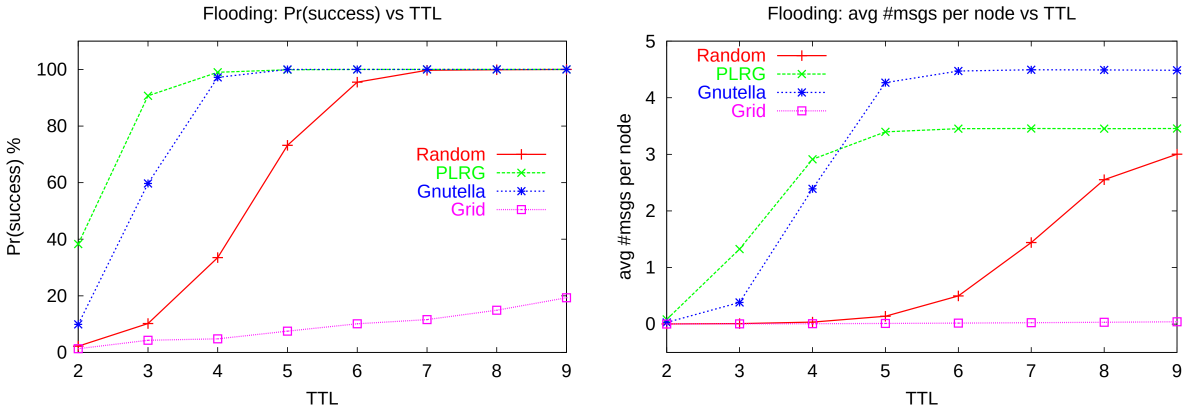


Figure 2: Pr(success) and per-node message overhead under various TTLs.

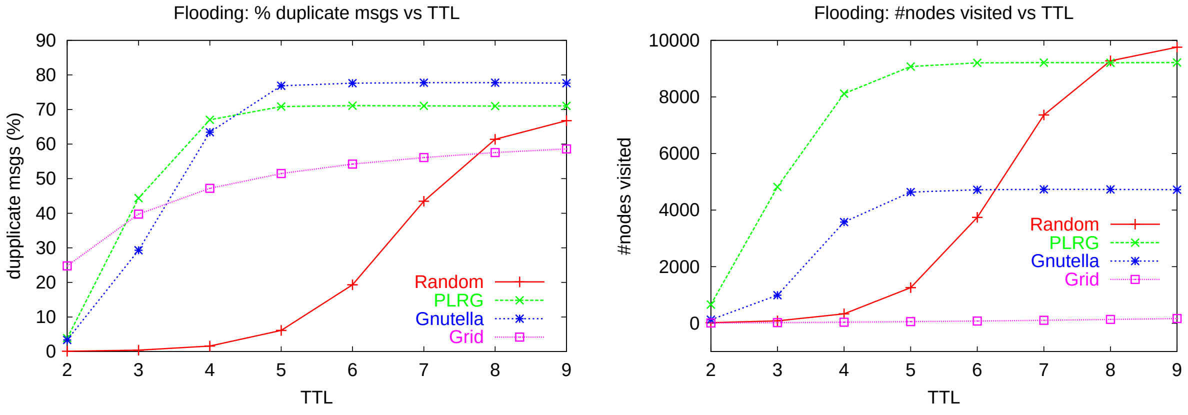


Figure 3: Message duplication percentage and node coverage under various TTLs.

hood that the next node already received the query) is the same as the fraction of nodes visited so far, as long as that fraction is small.

The random graph is also better for load distribution among its nodes. In the random graph, the maximum load on any one node is logarithmic to the total number of nodes that the search visits. In contrast, the high degree nodes in PLRG and Gnutella graphs have much higher load than other nodes. Due to space constraints we omit the data on peak # of messages here.

4. FINDING BETTER SEARCH METHODS

Our first try is to address the TTL selection problem.

4.1 Expanding Ring

One might suggest asking nodes to check with the original requester before forwarding the query to neighbors. However, this approach can lead to message implosion at the requester node. Hence, we do not adopt this approach.

Instead, we can use successive floods with increasing TTLs. A node starts a flood with small TTL, If the search is not successful, the node increases the TTL and starts another flood. The process repeats until the object is found. We expect this method to perform particularly well when hot objects are replicated more widely than cold objects, which is likely the case in practice. This method is called “expanding ring.”

To understand how well expanding ring works, we measure the average stopping TTL for searches to objects with increasing replication ratios. In our simulations the expanding ring starts with TTL=1, and expands the TTL linearly by 2 each time. Figure 4 shows the results for various topologies. As we can see, expanding ring successfully reins in the TTL as the object's replication ratio increases. While searches for objects with low replication ratio need TTLs larger than 5, the searches stop at TTL of 1 or 2 when the object replication is over 10%.

We also record the average number of messages a node has to process (Figure 4). Comparing the message overhead between flooding and expanding ring, we can see that, for objects that are replicated at 0.125% of the nodes, even if flooding uses the best TTL for each network topology, expanding ring still halves the per-node message overhead.

To understand the overall impact of expanding rings in a P2P network, we run a large set of simulations as explained in section 2.3. The results are shown in the “expanding ring” column in Tables 2 through 5.

The results show that, despite the successive retries, expanding ring still reduces message overhead significantly compared with regular flooding with a fixed TTL. The savings are obtained across all query and replication distributions.

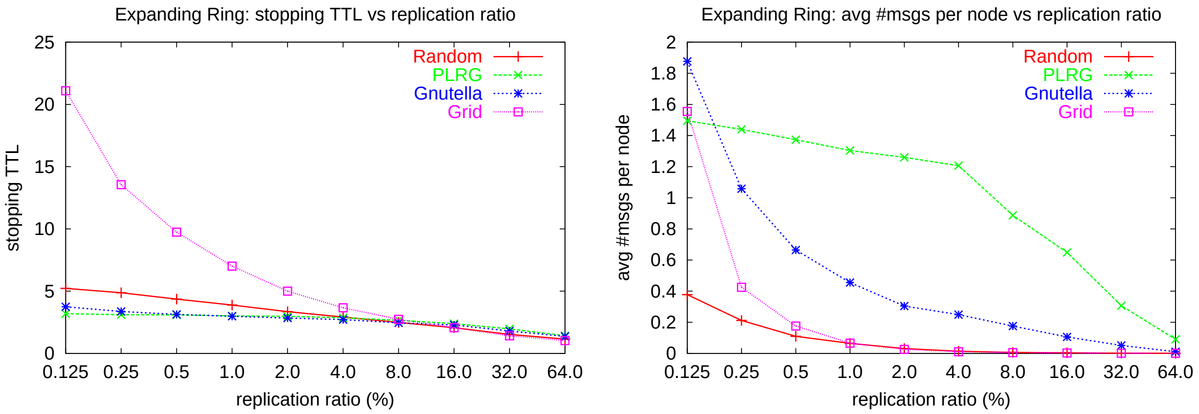


Figure 4: stopping TTL and per-node message overhead for expanding ring under various replication ratios.

The improvements are also more pronounced for Random and Gnutella graphs than for the PLRG graph, partly because the very high degree nodes in PLRG graph reduce the opportunity for incremental retries in expanding ring.

Expanding ring achieves the savings at the expense of slight increase in the delays to find the object. As we can see, for Random, PLRG and Gnutella, the average number of hops only increases from 2 to 4 in flooding to 3 to 6 in expanding ring, which we believe is tolerable for users.

Though expanding ring solves the TTL selection problem, it does not address the message duplication issue inherent in flooding. Inspection of simulation results shows that the duplication contributes significantly to the message overhead, particularly for PLRG and Gnutella graphs. To reduce message duplication, we try a different approach, random walk.

4.2 Random Walks

Random walk is a well-known technique, which forwards a query message to a randomly chosen neighbor at each step until the object is found. We call this message a "walker".

The standard random walk (which uses only one walker) can cut down the message overhead by an order of magnitude compared to expanding ring across the network topologies. However, there is also an order of magnitude increase in user-perceived delay.

To reduce the delay we increase the number of "walkers.". That is, instead of just sending out one query message, a requesting node sends k query messages, and each query message takes its own random walk. The expectation is that k walkers after T steps should reach roughly the same number of nodes as 1 walker after kT steps, and indeed simulations confirm that. Therefore, by using k walkers, we can expect to cut the delay down by a factor of k.

We experimented with different number of walkers. With more walkers, we can find objects faster, but also generate more loads. And when the number of walkers is big enough, increasing it further yield little reduction in the number of hops, but significantly increases the message traffic. Usually, 16 to 64 walkers give good results.

Since multiple-walker random walks require a mechanism to terminate the walks, we experimented with two methods, TTL and "checking.". TTL means that, similar to flooding, each random walk terminates after a certain number of hops. "Checking" means that a walker periodically checks with the original requester before walking to the next node. The checking method still uses a TTL, but the TTL is very large and is mainly used to prevent loops.

Our simulations show that checking is the right approach for terminating searches in random walks. The TTL approach runs into the same TTL selection issue in flooding. Meanwhile, since there are a fixed number of walkers (typically 16 to 64), having the walkers check back with the requester will not lead to message implosion at the requester node. Of course, checking does have overhead; each check requires a message exchange between a node and the requester node. Further experiments show that checking once at every fourth step along the way strikes a good balance between the overhead of the checking messages and the benefits of checking.

Tables 2 through 5 compare the discussed search methods under all combinations of query and replication distributions for the four network topologies. There are two sets of columns in each table; the first set are results of the queries to the top 50% hottest objects, and the second set are results of all queries.

The results show that, compared to flooding, the 32-walker random walk reduces message overhead by roughly *two orders of magnitude* for all queries across all network topologies, at the expense of slight increase in the number of hops (increasing from 2-6 to 4-15). The 32-walker random walk generally outperforms expanding ring as well, particularly in PLRG and Gnutella graphs.

We also studied an improvement to the above approach by asking each node to keep state. Each query has a unique ID and all its k walkers are tagged with that ID. For each ID, a node remembers the neighbors to which it has forwarded queries of that ID, and when a new query with the same ID arrives, the node forwards it to a different neighbor (randomly chosen). This state keeping accelerates the walks because walkers are less likely to cover the same route and hence they visit more nodes. Simulation results, also shown in the tables, confirm the improvement. Compared with random walks without state keeping, random walk with state keeping shows the biggest improvement in Random and Grid graphs, reducing message overhead by up to 30%, and reducing number of hops by up to 30%. However, the improvements for PLRG and Gnutella graphs are small. Hence, de-

distribution model		50 % (queries for hot objects)				100 % (all queries)			
query/replication	metrics	flood	ring	check	state	flood	ring	check	state
Uniform / Uniform	#hops	3.40	5.77	10.30	7.00	3.40	5.77	10.30	7.00
	#msgs per node	2.509	0.062	0.031	0.024	2.509	0.061	0.031	0.024
	#nodes visited	9220	536	149	163	9220	536	149	163
	peak #msgs	6.37	0.26	0.22	0.19	6.37	0.26	0.22	0.19
Zipf-like / Proportional	#hops	1.60	2.08	1.72	1.64	2.51	4.03	9.12	6.66
	#msgs per node	1.265	0.004	0.010	0.010	1.863	0.053	0.027	0.022
	#nodes visited	6515	36	33	47	7847	396	132	150
	peak #msgs	4.01	0.02	0.11	0.10	5.23	0.20	0.17	0.14
Zipf-like / Square root	#hops	2.23	3.19	2.82	2.51	2.70	4.24	5.74	4.43
	#msgs per node	2.154	0.010	0.014	0.013	2.308	0.031	0.021	0.018
	#nodes visited	8780	92	50	69	8983	269	89	109
	peak #msgs	5.88	0.04	0.16	0.16	6.14	0.12	0.17	0.16

Table 2: Static simulation results for Random Graph. The first set of columns are results of queries to the top 50% most popular objects; the second set of columns are results of all queries. "flood" is flooding with TTL=8. "ring" is expanding ring. "walk" is 32-walker random walk with checking. "state" is 32-walker random walk with checking and state keeping. #objects m is 100, average replication ratio is 1.0%, and the parameter α in the Zipf-like query distribution is 1.20.

distribution model		50 % (queries for hot objects)				100 % (all queries)			
query/replication	metrics	flood	ring	check	state	flood	ring	check	state
Uniform / Uniform	#hops	2.37	3.50	8.95	8.47	2.37	3.50	8.95	8.47
	#msgs per node	3.331	1.325	0.030	0.029	3.331	1.325	0.030	0.029
	#nodes visited	8935	4874	147	158	8935	4874	147	158
	peak #msgs	510.4	132.7	12.3	11.7	510.4	132.7	12.3	11.7
Zipf-like / Proportional	#hops	1.74	2.36	1.81	1.82	2.07	2.93	9.85	8.98
	#msgs per node	2.397	0.593	0.011	0.011	2.850	0.961	0.031	0.029
	#nodes visited	6969	2432	43	49	7923	3631	136	145
	peak #msgs	412.7	58.3	4.9	5.1	464.3	98.9	12.7	11.7
Zipf-like / Square root	#hops	2.07	2.94	2.65	2.49	2.21	3.17	5.37	4.79
	#msgs per node	3.079	0.967	0.014	0.014	3.199	1.115	0.021	0.020
	#nodes visited	8434	3750	62	69	8674	4200	97	103
	peak #msgs	496.0	93.7	6.3	6.3	499.6	111.7	8.9	8.4

Table 3: Static simulation results for Power-Law Random Graph (PLRG). The legends are the same as in Table 2, except that the TTL for "flood" is 5.

distribution model		50 % (queries for hot objects)				100 % (all queries)			
query/replication	metrics	flood	ring	check	state	flood	ring	check	state
Uniform / Uniform	#hops	2.39	3.40	7.30	6.11	2.39	3.40	7.30	6.11
	#msgs per node	4.162	0.369	0.051	0.045	4.162	0.369	0.051	0.045
	#nodes visited	4556	933	141	151	4556	933	141	151
	peak #msgs	64.9	6.4	1.3	1.2	64.9	6.4	1.3	1.2
Zipf-like / Proportional	#hops	1.60	2.18	1.66	1.66	2.03	3.05	9.39	7.94
	#msgs per node	2.961	0.109	0.021	0.021	3.548	0.423	0.058	0.051
	#nodes visited	3725	357	49	60	4137	810	143	153
	peak #msgs	43.8	2.0	0.7	0.8	54.5	7.0	1.6	1.5
Zipf-like / Square root	#hops	1.88	2.70	2.31	2.15	2.10	3.02	4.61	4.12
	#msgs per node	3.874	0.208	0.027	0.026	4.007	0.302	0.038	0.035
	#nodes visited	4404	621	67	80	4479	789	101	114
	peak #msgs	62.5	3.8	0.8	0.9	63.8	5.3	1.1	1.1

Table 4: Static simulation results for Gnutella Graph. The legends are the same as in Table 2, except that the TTL for "flood" is 5.

distribution model		50 % (queries for hot objects)				100 % (all queries)			
query/replication	metrics	flood	ring	check	state	flood	ring	check	state
Uniform / Uniform	#hops	6.52	19.15	27.95	15.20	6.52	19.15	27.95	15.20
	#msgs per node	0.472	0.070	0.068	0.041	0.472	0.070	0.068	0.041
	#nodes visited	1692	128	107	128	1692	128	107	128
	peak #msgs	0.72	0.18	0.30	0.16	0.72	0.18	0.30	0.16
Zipf-like / Proportional	#hops	1.70	2.32	1.95	1.77	4.71	19.04	33.78	15.14
	#msgs per node	0.321	0.003	0.011	0.010	0.392	0.120	0.082	0.040
	#nodes visited	1398	14	22	28	1533	118	111	121
	peak #msgs	0.57	0.02	0.07	0.06	0.64	0.25	0.26	0.16
Zipf-like / Square root	#hops	2.77	4.64	4.60	3.32	4.31	10.66	15.53	8.22
	#msgs per node	0.437	0.008	0.018	0.015	0.450	0.034	0.041	0.025
	#nodes visited	1647	31	34	42	1656	70	67	77
	peak #msgs	0.68	0.04	0.10	0.08	0.68	0.10	0.19	0.12

Table 5: Static simulation results for Grid Graph. The legends are the same as in Table 2, except that the TTL for "flood" is 32.

pending on the implementation overhead of state keeping, each P2P network should decide separately whether state keeping is worthwhile.

4.3 Principles of Scalable Search in Unstructured Networks

Our results show that the k-walker random walk is a much more scalable search method than flooding. However, perhaps more important than this conclusion is the understanding we have gained from this exercise. The key to scalable searches in unstructured network is to cover the right number of nodes as quickly as possible and with as little overhead as possible. In unstructured network, the only way to find objects is to visit enough nodes so that, statistically speaking, one of the nodes has the object. However, in reaching the required node coverage, one must pay attention to the following:

- *Adaptive termination is very important.* TTL-based mechanism does not work. Any adaptive/dynamic termination mechanism must avoid message implosion at the requester node. The checking method described above is a good example of adaptive termination.

- *Message duplication should be minimized.* Preferably, each query should visit a node just once. More visits are wasteful in terms of the message overhead.

- *Granularity of the coverage should be small.* Each additional step in the search should not significantly increase the number of nodes visited. This is perhaps the fundamental difference between flooding and multiple-walker random walk. In flooding, an additional step could exponentially increase the number of nodes visited; in random walk, it increases only by a constant.

5. REPLICATION: THEORY

Our study in the previous section examined how one should search for an object, assuming that it is replicated at some random locations in the network. In certain P2P systems such as Gnutella, only nodes that request an object make copies of the object. Other P2P systems such as Freenet allow for more proactive replications of objects, where an object may be replicated at a node even though the node has not requested the object. For such systems, we must answer the question: how many copies of each object should there be so that the search overhead for the object is minimized, assuming that the total amount of storage for objects in the network is fixed? Answers to this question have implications to non-proactive replication systems as well, because the information of an object's location could be proactively replicated to expedite the searches.

This question has been addressed in great theoretical detail in [8]. Here we review those results and evaluate the resulting algorithms via simulation. Consider the simple model used in [8] where there are m objects and n sites(nodes). Each object i is replicated at r_i random (distinct) sites (recall that $R = \sum_i r_i$), and that object i is requested with relative rates q_i, where we normalize this by setting $\sum_i q_i = 1$. For convenience, we assume that query and replication strategies are such that $1 \ll r_i \leq n$ and that searches go on until a copy is found. The other cases are dealt with in [8], and the conclusions are consistent with, but a bit messier than, what we present here. Search consists of randomly probing sites until the desired object is found. Thus, the probability $Pr(k)$ that the object is found on the k'th probe is given by: $Pr_i(k) = \frac{r_i}{n}(1 - \frac{r_i}{n})^{k-1}$. An object's *average search size* A_i is merely the inverse of the fraction of sites which have replicas of the object: $A_i = \frac{n}{r_i}$. We are interested in the average search size of all the objects: $A = \sum_i q_i A_i = n \sum_i \frac{q_i}{r_i}$. This metric essentially captures the message overhead of efficient searches.

If there were no limit on r_i then clearly the optimal strategy would be to replicate everything everywhere, and all searches become trivial. Instead, we assume that the average number of these replicas per site, $\rho = \frac{R}{n}$, is fixed and less than m. The question is how to allocate these R replicas among the m objects.

The simplest replication strategy is to create the same number of replicas of each object: $r_i = \frac{R}{m}$. We call this *uniform* replication strategy. In this case the average search size $A_{uniform}$ is given by:

$$A_{uniform} = \sum_i q_i \frac{m}{\rho} = \frac{m}{\rho}$$

which is independent of the query distribution.

It is clear that uniformly replicating all objects, even those that are not frequently queried, is inefficient. A more natural policy is to replicate *proportional* to the querying rate: $r_i = Rq_i$. This should reduce the search sizes for the more popular objects. However, a quick calculation reveals that the *average* remains the same:

$$A_{proportional} = n \sum_i \frac{q_i}{Rq_i} = \frac{m}{\rho} = A_{uniform}$$

Thus, the Proportional and Uniform replication strategies yield *exactly* the same average search size, and that average search size is independent of the query distribution.

Another important metric that captures the load balancing ability of a replication strategy is *utilization rate*, $U_i = R\frac{q_i}{r_i}$. That is, the rate of requests that a replica of object i serves (the random probing search process implies that all replicas of the same object have the same utilization rate). Note that the average utilization over all objects $U = \sum_i r_i U_i / R = 1$ is *fixed* for all replication strategies. The maximum utilization $\max_i U_i$, however, varies considerably.

The distributions of average search sizes and utilization rates for an object are quite different between the Uniform and Proportional strategies. For Uniform replication, all objects have the same average search size, but replicas have utilization rates *proportional* to their query rates. Proportional replication achieves perfect load balancing with all replicas having the same utilization rate, but average search sizes vary with more popular objects having smaller average search sizes than less popular ones. Objects whose query rates are greater than average (i.e., greater than $\frac{1}{m}$) do better with Proportional replication, and the other objects do better with Uniform replication. Interestingly, the weighted average of the search sizes over all objects balances out to be unchanged.

Square-Root Replication Given that Uniform and Proportional have the same average search size, a natural question is what is the optimal way to allocate the replicas so that the average search size is minimized? A simple calculation (see [8]) reveals that Square-Root replication is optimal; that is, A is minimized when $r_i = \lambda\sqrt{q_i}$ where $\lambda = \frac{R}{\sum_i \sqrt{q_i}}$. The average search size is

$$A_{optimal} = \frac{1}{\rho}(\sum_i \sqrt{q_i})^2$$

Table 6 lists properties of the three replication strategies. Square-Root replication is such that both average search size and utilization rate vary per object, but the variance in utilization is considerably smaller than with Uniform, and the variance in average search size is considerably smaller than with Proportional.

	Uniform	Proportional	Square-Root
A	$\rho^{-1}m$	$\rho^{-1}m$	$\rho^{-1}(\sum_i \sqrt{q_i})^2$
r_i	R/m	$q_i R$	$R\sqrt{q_i}/\sum_j \sqrt{q_j}$
$A_i = n/r_i$	$\rho^{-1}m$	$(\rho q_i)^{-1}$	$\rho^{-1}\sum_j \sqrt{q_j}/\sqrt{q_i}$
$U_i = Rq_i/r_i$	$q_i m$	1	$\sqrt{q_i}\sum_j \sqrt{q_j}$

Table 6: Comparing the three replication strategies

5.1 Truncated Pareto Distribution

Assuming $q_1 \geq q_2 \geq \cdots \geq q_m$, let $R_{m,\alpha}$ be a truncation to m objects of a Pareto distribution with shape parameter α. Thus, $q_i = i^{-\alpha-1}/B_{\alpha+1,m}$ $(i = 1, \ldots, m)$, where $B_{\alpha+1,m}$ is the normalization factor. With Square-Root replication we obtain $r_i/R = i^{(-\alpha-1)/2}/B_{(\alpha+1)/2,m}$, average search size

$$\begin{aligned} \rho A_{optimal} &= (B_{(\alpha+1)/2,m})^2/B_{\alpha+1,m} \\ &\approx \begin{cases} 4\alpha m^{1-\alpha}/(1-\alpha)^2 & (\alpha < 1) \\ \ln^2 m & (\alpha = 1) \\ 4\alpha/(\alpha-1)^2 & (\alpha > 1) \end{cases} \end{aligned}$$

We then compute *gain factor*, $A_{uniform}/A_{optimal}$, of using Square-Root rather than Uniform or Proportional replication. For Pareto distribution with $\alpha > 1$, the optimal average search size is *constant*. Thus, the gain factor is $\Theta(m)$. The gain factor is $\Theta(m^\alpha)$ for Pareto with $\alpha < 1$ and $\theta(m/\ln^2 m)$ for $\alpha = 1$.

Figure 5 helps visualize the different properties of the three replication strategies. Both Uniform and Square-Root allocate to popular objects less than their "fair share" and to less popular objects more than their "fair share" of replicas, but Square-Root does so to a lesser extent. The variance in average search sizes of different objects with Square-Root is considerably smaller than with Proportional. The maximum utilization rate with Square-Root, is much lower than with Uniform (although larger than Proportional which provides optimal load balancing).

5.2 Achieving Square-Root Replication

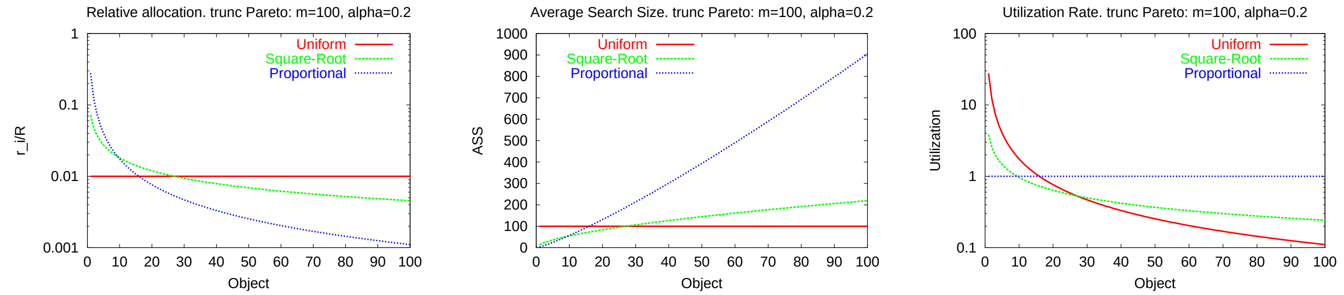
Assume that each query keeps track of the search size, each time a query is finished, the object is copied to a number of sites proportional to the number of probes. This means that on average the i'th object will be replicated $c\frac{n}{r_i}$ times each time a query is issued (where c is an arbitrary constant). Thus, the number of copies r_i can be roughly described by the differential equation $\dot{r_i} = q_i c\frac{n}{r_i}$, where $\dot{r_i}$ is the time derivative of r_i.

If we look at the ratio of two objects, ask how the logarithm of this quantity changes, we find that, setting $z_{i,j} = \ln\frac{r_i}{r_j}$, we have $\dot{z} = cn(\frac{q_j}{r_j^2} - \frac{q_i}{r_i^2})$. Thus, Square-Root replication, $r_i = \lambda\sqrt{q_i}$ is a fixed point of this equation. This heuristic calculation suggests that perhaps replicating proportional to the number of sites probed would yield Square-Root replication.

Our analysis above makes some implicit assumptions on the the *deletion of replicas*, that is, replicas disappear over time and new replicas are created. The steady state is achieved when the creation rate equals the deletion rate. To achieve this, the lifetimes of replicas must be *independent* of object identity or query rate. Examples of deletion processes that have this independence are: assigning fixed lifetimes (or lifetimes from a fixed distribution) for each replica, subject replicas at each site to First In First out (FIFO) replacement, or perform random deletions. Usage-based replacement policies such as Least Recently Used (LRU) or Least Frequently Used (LFU) *do not* have this independence property. These policies could *impede* the Square-Root scheme:

Note that unlike Freenet's replication algorithm, the replication strategies studied here do not attempt to cluster certain group of objects in certain regions of the network. In other words, they do not produce any correlation between routing and object locations, or, "structure," in the P2P network.

6. REPLICATION: EVALUATION



Relative allocation r_i/R Average search size ρA_i Utilization rate U_i

Figure 5: Uniform, Proportional, and Square-Root strategies on a truncated Pareto distribution

We observe that there are two replication strategies that are easily implementable. One is "owner replication", where, when a search is successful, the object is stored at the requester node only. The other is "path replication", where, when a search succeeds, the object is stored at all nodes along the path from the requester node to the provider node. Owner replication is used in systems such as Gnutella. Path replication is used in systems such as Freenet.

The analysis in the previous section suggests that square-root replication distribution is needed to minimize the overall search traffic, and an object should be replicated at the number of nodes that is proportional to the number of probes that the search required. If a P2P system uses the k-walker random walk as the search algorithm, then on average, the number of nodes between the requester node and the provider node is $1/k$ of the total nodes visited. Path replication in this system should result in square-root distribution.

However, path replication tends to replicate objects to nodes that are *topologically* along the same path. To understand how this impacts the overall search traffic, we also study a third replication algorithm, "random replication." In random replication, once a search succeeds, we count the number of nodes on the path between the requester and the provider, p, then randomly pick p of the nodes that the k walkers visited to replicate the object. "Random replication" is harder to implement, but the performance difference between it and path replication highlights the topological impact of path replication.

We design a set of dynamic simulations to study the performance of the three replication strategies in the Random graph network topology. A simulation starts by placing the m distinct objects randomly into the network. Then the *Query Generator* starts to generate queries according to a Poisson process with average generating rate at 5 queries per second. The query distribution among the m objects follows Zipf-like distribution with a given α value. The α value for the results presented here is 1.20. (We also ran simulations with $\alpha = 0.80$ and $\alpha = 2.40$. The results are similar.) For each query, a node (that doesn't have the requested object yet) is randomly chosen to start the query. For the search method, we use the 32-walker random walk with state keeping, with checking at every fourth step. Each node can store at most *objAllow* objects (40 in our simulations). The replacement policy is Random Deletion. "Snapshots" are taken for every 2,000-query chunks. To allow for enough "warming up" process, we run each simulation for 10,000 seconds, and look at the later part of the simulation.

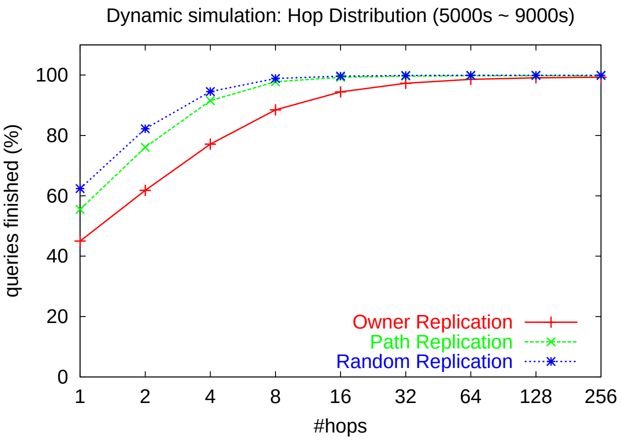


Figure 7: Cumulative distribution of the number of hops under the three replication strategies.

For each replication strategy, we examine:

- what kind of replication ratio distribution does the strategy generate?
- what is the average number of messages per node in a system using the strategy?
- what is the distribution of number of hops in a system using the strategy?

Figure 6 shows log-log plots of the distribution of replication ratios under path replication and random replication. We also plot the distribution that is the square root of the query distribution. Confirming our theoretical predictions, the results show clearly that both path replication and random replication generates replication ratios that are quite close to square-root of query ratios.

Table 7 lists the average number of messages a node has to process during the simulation. The result shows clearly the benefit of square-root distribution on reducing search traffic. Path replication and random replication reduces the overall message traffic by a factor of three to four. Hence, proactive replication such as path or random replication can improve the scalability of P2P systems significantly.

Much of the traffic reduction comes from reducing the number of hops it takes to find an object. Figure 7 shows the cumulative hop distribution for all queries under the

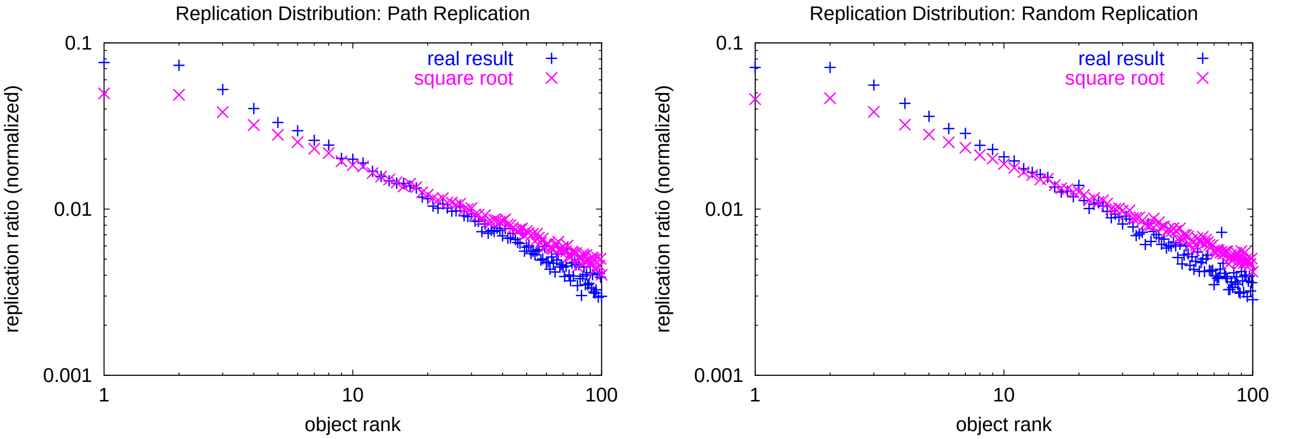


Figure 6: Distribution of replication ratios under Path replication and Random replication.

	Owner Replication	Path Replication	Random Replication
avg #msgs per node	56542.6	19155.5	14463.0
factor of improvement	1	2.95	3.91

Table 7: Message traffic of different replication strategies.

three replication strategies. Path replication and random replication clearly outperform the owner replication; for example, the percentage of queries that finish within four hops are 71% for owner replication, 86% for path replication and 91% for random replication.

The results also show that random replication improves upon the path replication. Thus, the topological effects of replicating along the path do hurt performance somewhat. Hence, if the implementation is not overly complex, a P2P system should adopt random replication instead of path replication.

7. RELATED WORK

As mentioned in the Introduction, there are several different kinds of P2P networks. The highly structured P2P networks, such as CAN [18], Chord [23], Past [20], and Tapestry [26], all use precise placement algorithms and specific routing protocols to make searching efficient. However these systems have not been widely deployed, and their ability to operate with extremely unreliable nodes has not yet been demonstrated. Moreover, they cannot deal with partial-match queries.

There are also many loosely structured P2P networks. Freenet [11], Free Haven [10], Mojo Nation [15] are but a few examples of this rapidly growing list. Many of them are storage systems that can protect the anonymity of both authors and readers. Some systems, such as Free Haven and Mojo Nation, focus on the trust, reputation management and security issues in Peer-to-Peer systems. Most of these loosely structured P2P networks use either directories or placement hints to improve the scalability of the search process. However, centralized directories don't scale well and placement hints don't handle partial-match queries.

An interesting paper by Adamic *et al.* [1] studies random-walk search strategies in power-law networks, and finds that by modifying walkers to seek out high degree nodes the search performance can be greatly increased. However, such strategies greatly reduce the scalability of the search algorithm, which is our focus and not the subject of [1], because then almost all queries are sent to the very high degree nodes, making them bear the burden of almost the entire query load of the network.

The random walk search style is used in Freenet as well. There, the walk is guided; each node uses hints to help it choose which node to forward the query to. It also uses only one "walker". [9] also studies ways to provide hints by building routing indices. In comparison, our focus is on unstructured network, where hints are not available.

Several search techniques for unstructured networks are discussed in [25]: iterative deepening, directed BFS and local indices. The first scheme is similar to the expanding ring scheme we examine here. Their work focuses on search techniques only and does not study replication strategies.

We found a wealth of information on Gnutella at web sites such as www.openP2P.com and gnutella.wego.com. We are also aware of a number of published research studies on the Gnutella network. For example, the freeloader phenomenon is examined in [2], and the topology and query distribution are studied in [21, 14]. However, none of these papers address the issue of better search algorithms or replication algorithms.

8. CONCLUSIONS AND FUTURE WORK

This paper reports our simulation and modeling studies of several search algorithms and replication strategies for decentralized, unstructured peer-to-peer networks.

From simulation studies, we have learned that scalable search algorithm designs for such networks should consider three properties: adaptive termination, minimizing message duplication, and small granularity of coverage. The flooding algorithm being used in Gnutella does not satisfy any of the properties. We show that it generates a lot of network traffic and does not scale well. The expanding ring approach improves the flooding algorithm by using an adaptive termination mechanism. It can find data reasonably quickly while reducing the network traffic substantially, sometimes by an order of magnitude. The k-walker random walk with checking approach can find data more quickly while reducing the traffic further by another order of magnitude, because it reduces the granularity of coverage by using a fixed number of random walkers.

We reviewed the calculations in [8] that show that, for a fixed average number of replicas per node, square-root replication distribution is theoretically optimal in terms of minimizing the overall search traffic. The simulations presented here demonstrate the effectiveness of the square-root replication policy and illustrated how it can be achieved in practice. We simulated owner, path and random replications, with the k-walker random walk with state keeping. Since path and random replications lead to square-root replication distribution, their overall message traffic is about four times less than the owner replication approach.

We have also learned from our simulation studies that uniformly random graphs are better for searching and data replication. The high degree nodes in power-law random graph and the current Gnutella network bear much higher load than average and introduce more duplication overhead in searches. The results imply that it is better to use P2P network building algorithms that form a uniformly random network topology.

This study is our first step towards understanding the properties of scalable search algorithms, replication strategies, and network topologies for decentralized, unstructured peer-to-peer networks. There are still many open issues to study. It would be useful to model various search algorithms with certain network topologies and study them analytically. The k-walker random walk with checking and state keeping has a lot of rooms to improve. Also, for the various replication strategies, we show that Square-Root replication minimizes the average search size, while Proportional replication achieves optimal load balance. It would be interesting to study the tradeoff between the two metrics, and see how the tradeoff can be achieved in a distributed system.

9. REFERENCES

[1] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power law networks. In *Phys. Rev. E64*, pages 46135–46143, 2001.

[2] E. Adar and B. A. Huberman. Free riding on gnutella. In *First Monday, http://www.firstmonday.dk/issues/ issue5_10/adar/index.html*, Oct. 2000.

[3] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of STOC'00*, pages 171–180, 2000.

[4] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing reference locality in the www. In *Proceedings of 1996 International Conference on parallel and Distributed Information Systems (PDIS '96)*, Dec. 1996.

[5] K. Calvert and E. W. Zegura. Gt-itm: Georgia tech internetwork topology models. In *http://www.cc.gatech.edu/projects/gtitm/*, 1997.

[6] Clip2.com. The gnutella protocol specification v0.4. In *http://www9.limewire.com/developer/ gnutella_protocol_0.4.pdf*, 2000.

[7] Clip2.com. Gnutella: To the bandwidth barrier and beyond. In *http://www.clip2.com/gnutella.html*, 2000.

[8] E. Cohen and S. Shenker. Replication strategies in unstructured peer-to-peer networks. In *Proceedings of ACM SIGCOMM'02*, Aug. 2002.

[9] A. Crespo and H. Garcia-Molina. Routing indices for peer-to-peer systems. In *Proceedings of the 22nd IEEE International Conference on Distributed Computing Systems (ICDCS)*, July 2002.

[10] *Free Haven website*. http://www.freehaven.net.

[11] *Freenet website*. http://freenet.sourceforge.net.

[12] D. Gallagher and R. Wilkerson. Network performance statistics for university of south carolina. In *http://eddie.csd.sc.edu*, Oct. 2001.

[13] *Gnutella website*. http://gnutella.wego.com.

[14] M. A. Jovanovic, F. S. Annexstein, and K. A. Berman. Scalability issues in large peer-to-peer networks - a case study of gnutella. Technical Report http://www.ececs.uc.edu/~mjovanov/Research/ paper.html, University of Cincinnati, 2001.

[15] *Mojo Nation*, 2001. http://www.mojonation.net.

[16] *Napster website*. http://www.napster.com.

[17] D. Plonka. Uw-madison napster traffic measurement. In *http://net.doit.wisc.edu/data/Napster*, Mar. 2000.

[18] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. In *Proceedings of ACM SIGCOMM'01*, Aug. 2001.

[19] J. Ritter. Why gnutella can't scale. no, really. In *http://www.darkridge.com/~jpr5/doc/ gnutella.html*, 2001.

[20] A. Rowstron and P. Druschel. Storage management and caching in past, a large-scale, persistent peer-to-peer storage utility. In *Proceedings of SOSP'01*, 2001.

[21] K. Sripanidkulchai. The popularity of gnutella queries and its implications on scalability. In *O'Reilly's www.openp2p.com*, Feb. 2001.

[22] S. D. G. Stefan Saroiu, P. Krishna Gummadi. A measurement study of peer-to-peer file sharing systems. Technical Report UW-CSE-01-06-02, Department of Computer Science & Engineering, University of Washington, 2002.

[23] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of ACM SIGCOMM'01*, Aug. 2001.

[24] K. Truelove. Gnutella: Alive, well, and changing fast. In *http://www.openp2p.com/pub/a/p2p/2001/01/25/ truelove0101.html*, Jan. 2001.

[25] B. Yang and H. Garcia-Molina. Efficient search in peer-to-peer networks. In *Proceedings of the 22nd IEEE International Conference on Distributed Computing Systems (ICDCS)*, July 2002.

[26] B. Y. Zhao, J. Kubiatowicz, and A. Joseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical Report UCB/CSD-01-1141, University of California at Berkeley, Computer Science Department, 2001.