Efficient Lookup on Unstructured Topologies*

Ruggero Morselli†, Bobby Bhattacharjee†,‡, Michael A. Marsh‡, and Aravind Srinivasan†,‡

Abstract

We present LMS, a protocol for efficient lookup on unstructured networks. Our protocol uses a virtual namespace *without imposing specific topologies*. It is more efficient than existing lookup protocols for unstructured networks, and thus is an attractive alternative for applications in which the topology cannot be structured as a Distributed Hash Table (DHT).

We present analytic bounds for the worst-case performance of our protocol. Through detailed simulations (with up to 100,000 nodes), we show that the actual performance on realistic topologies is significantly better. We also show in both simulations and a complete implementation (which includes over five hundred nodes) that our protocol is inherently robust against multiple node failures and can adapt its replication strategy to optimize searches according to a specific heuristic. Moreover, the simulation demonstrates the resilience of LMS to high node turnover rates, and that it can easily adapt to orders of magnitude changes in network size. The overhead incurred by LMS is small, and its performance approaches that of DHTs on networks of similar size.

1 Introduction

Large peer-to-peer networks require efficient object lookup mechanisms. Flooding searches, while adequate for small networks, quickly become untenable as networks grow larger. Consequently, research has progressed along two avenues: more efficient lookup protocols that operate on general (unstructured) topologies, and fundamentally new lookup strategies that achieve bounded worst-case performance by adding constraints to the network topology. The latter, while very efficient, might not be usable when the peer-to-peer application places its own constraints on the topology. This is especially the case when the links in the topology have particular significance, such as expressing statements of trust or belief between peers. From an algorithmic perspective, this sort of network topology appears to be unstructured.

We approach this issue by considering a model where a graph G is given (the topology), such that each peer in the system corresponds to a vertex u of G and can only send messages to peers that correspond to neighbors of u in G. Developing distributed protocols in this model is not only an interesting theoretical problem, but has several practical applications, such as that outlined in Section 6. In this paper, we present the Local Minima Search (LMS) protocol for unstructured topologies. LMS extends the notion of random walks, which to date have proved the most promising approach to improving search performance on unstructured networks [18, 14, 6]. In addition, LMS borrows the ideas of *namespace virtualization* and *consistent hashing* employed by constrained-topology protocols such as distributed hash tables (DHTs) [29, 33, 31, 19]. Namespace virtualization maps both peers and objects to identifiers in a single large space.

In LMS, the owner of each object places replicas of the object on several nodes. Like in a DHT, LMS places replicas onto nodes which have IDs "close" to the object. Unlike in a DHT, however, in an unstructured topology there is no mechanism to route to the node which is the closest to an item. Instead, we introduce

*Full version. Technical report CS-TR-4772, Department of Computer Science, University of Maryland. Obsoletes CS-TR-4593. An extended abstract of this paper appeared in ACM PODC 2005, Las Vegas, NV, July 2005. This material is based upon work supported in part by: NSF grant 0208005, ITR Award CNS-0426683, NSF grant CCR-0310499, NSF award ANI0092806 and DoD contract MDA90402C0428. We thank past referees of this work for their valuable comments.

†Department of Computer Science, University of Maryland, College Park, MD 20742.

‡Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742.

the notion of a *local minimum*: a node u is a local minimum for an object if and only if the ID of u is the closest to the item's ID in u's *neighborhood* (those nodes within h hops of u in the network, where h is a parameter of the protocol, typically 1 or 2). In general, for any object there are many local minima in a graph, and replicas are placed onto a subset of these. During a search, random walks are used to locate minima for a given object, and a search succeeds when a minimum holding a replica is located.

Let d_h be the minimum size of the h-hop neighborhood of any node in G and let g be the eigenvalue gap of a random walk in the graph (see Section 4). LMS achieves its performance by storing $O(\sqrt{n/d_h})$ replicas, and with a message complexity (in its lookups) of $O(\sqrt{n/d_h} \cdot (\log n)/g)$. g is a constant for a wide range of randomly-grown topologies. This is notably worse than DHTs, but is a considerable improvement over other (essentially linear-time) lookup techniques in networks that cannot support a structured protocol, and a vast improvement over flooding-based searches. Furthermore, as we shall see, LMS provides a high degree of fault-tolerance.

Note that the LMS protocol provides a *probabilistic* guarantee of success. Depending on the number of replicas placed and the number of search probes, an object may not be found even if it exists in the network, though the probability of failure can be made arbitrarily small. In Section 4, we derive expressions for the number of necessary replicas and probes for specific probabilities of success, for arbitrary graphs G. In particular, we deal with the following problem. Let D be an arbitrary distribution on a set of size k. Suppose we sample $r + s$ times independently from D; what is a good upper bound on the probability that the multiset of the first r samples, is disjoint from the last s samples? We present an upper-bound of $\exp(-\Omega(s \cdot \min\{r/k, 1\}))$, assuming without loss of generality that $s \leq r$. At first sight, this may appear related to the "birthday paradox", and D being the uniform distribution may appear to be the worst case. This turns out to not be true — this problem is more complex, as is illustrated in Section 4. Our proof approach also facilitates a proof of the fault-tolerance of LMS.

The object placement component of LMS distrubutes replicas randomly throughout the network. This means that even if the LMS lookup is not used (e.g. for multi-attribute searches when the object id-based lookup is not applicable) flooding will succeed with high probability even with relatively small bounded propagation. LMS can also be augmented with index-based searches in the same manner as DHTs. This flexibility suggests that LMS could become the underlying platform of choice for building wide-area applications.

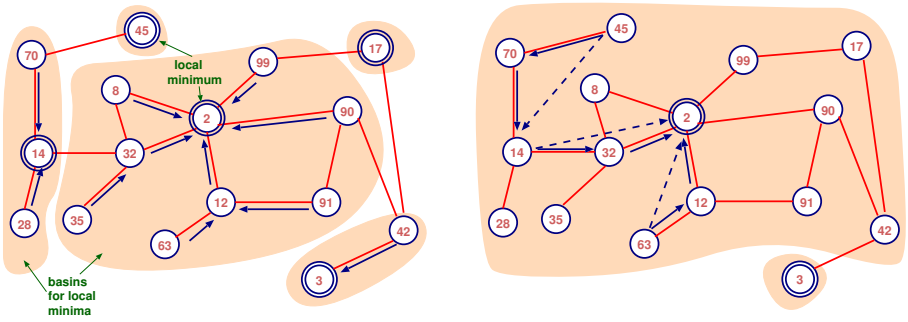
We start with a description of related work in Section 2. The primary contribution of this paper is the base LMS protocol described in Section 3, for which we derive expected and worst-case performance bounds in Section 4. A particularly novel feature of our analysis is that the lower bounds hold regardless of G. We also analyze, in Section 5, the performance of LMS over realistic topologies via a set of comprehensive simulations on networks with 10^5 peers and over a testbed with 512 peers. In Section 6, we finally conclude.

2 Related work

Gnutella [15] is probably the most-studied unstructured peer-to-peer network; much work has been done to understand Gnutella dynamics [28, 30, 7]. The original Gnutella search protocol was based on naive flooding. An improved flooding algorithm is discussed in [16]; it reduces the number of messages per search while still reaching the entire network. More sophisticated search techniques based on random walks are described in [18]. Their analysis yields a search cost of $\frac{n}{r}$, where n is the size of the network and r is the number of replicas, although it is based on the assumption that replicas are placed on uniformly random nodes, while the paper does not describe any placement protocol. As a comparison, LMS achieves a search cost of $O(\frac{n}{rd_h} \log n)$. In Section 5 we directly compare LMS to the best protocols described in [18].

The Gia system [6] improves on Gnutella using: topology adaptation, flow control and biased random walks. We do not compare LMS and Gia, because our model does not allow topology adaptation and, by design, the LMS substrate is not responsible for node heterogeneity and node congestion, without which flow control and biased random walks are of no use. Note that one can implement the latter two techniques on top of LMS.

Highly-efficient DHT lookup algorithms (e.g., [31, 29, 33, 25, 19]) impose structure on the topology in order to achieve their performance bounds. Beehive [24] is a replication framework that can be built on top



(a) 1-hop Neighborhoods

(b) 2-hop Neighborhoods

Figure 1: Local minima, basins and deterministic forwarding. Nodes are labelled with their distances from the key. Concentric circles denote local minima, and shaded regions their basins. Arrows indicate the path from a node towards its local minimum. (a) All forwarding arrows for $h = 1$. (b) Forwarding arrows for two nodes with $h = 2$. A dashed arrow indicates the target towards which a node forwards a probe when different than the next hop.

of a DHT. These systems employ the same namespace virtualization of which we make use. A virtualized namespace has previously been used on an unstructured topology in Freenet [8], though for a different reason.

In [2] the authors build a P2P system where publishing involves storing on s random nodes in the graph and searching involves querying s independently chosen random nodes. Unlike LMS, [2] requires that the nodes may modify the topology of the neighborhood graph. The authors prove a result that is similar to Theorem 4.1(i) in this paper, although with a different technique; we thank a referee for bringing this result to our attention. Other differences are: (a) LMS restricts replica placement to local minima, which improves efficiency significantly; (b) our Theorem 4.1 is generalized beyond the special case in [2]; and (c) our proof technique for Theorem 4.1 also helps us show that LMS is robust under the loss of some replicas.

YAPPERS [12] is a P2P system that, like LMS, assumes a "given topology" model and combines structured and unstructured designs. In YAPPERS, nodes publish an item by storing it at a node of the 2-hop neighborhood, which is structured as a small DHT. Search is achieved by flooding all and only the nodes in the network that *might* have the item, and requires nodes to keep state for an extended local neighborhood (within 5 hops). The authors do not present a formal analysis.

Related to the problem of lookup in a given topology is the work on name-independent routing (NIR) in a graph [4, 1, 3]. NIR allows a node to send a message to any other node, in a given topology setting. Existing work on NIR assumes that the topology graph never changes and requires an initial setup phase performed by a centralized algorithms that knows the entire graph.

Bloom filters [5, 21], compact digests of sets of elements, have previously been considered for structured topologies [26] as a way to provide fast lookup. In this paper, we present an application of Bloom filters on a completely unstructured topology. Theoretical properties of random walks have been exploited by [14] to improve search in unstructured networks.

3 LMS protocol description

We assume a system of principals (*nodes* or *peers*) structured as an overlay network on top of a communications infrastructure such as the Internet. The topology of this overlay network is dictated by external requirements, and its maintenance is beyond the scope of our protocol. We assume that nodes can commu-

nicate with one another if and only if they are neighbors in the overlay topology.[1]

Nodes in the system have unique identifiers generated uniformly at random from an *ID space* of all bit strings of some length λ. λ must be chosen large enough to guarantee uniqueness with high probability (for example, 160 bits). How this identifier is constructed is in general application-specific.[2]

We define a node v's *neighborhood* as the set of nodes at most h hops away from v in the overlay. For each of these nodes, v knows its unique identifier, how many hops away it is, and v's next hop towards it. We do not present a protocol for propagating this information, but it is easily implemented by repeated exchanges of a node's known $h-1$-hop neighborhood. Note that nodes have no knowledge of the topology beyond this neighborhood.

3.1 Protocol overview

LMS enables nodes to publish data items (i.e. files, documents) and to retrieve data items published by others. Items published by the system are given *key identifiers* (or simply *keys*) in the same ID space as the nodes, for example by hashing the name or contents of the item. The protocol then attempts to store an item at nodes that are close to its key in the (circular) ID space. In terms of the distance between a node and key, we do not find the global minimum (as in a distributed hash table), but rather a *local* minimum.

Publishing an item involves storing *replicas* of the item at a number of randomly selected local minima; retrieving an item involves querying randomly selected local minima until one is found that holds a replica. Crucially, the distribution of these random selections is typically not the uniform distribution; it is a function of the underlying topology, and is naturally generated by the distributed algorithms that we employ. Intuitively, the more replicas are placed, the easier it will be to find one of them. Random minimum selection is accomplished by performing a random walk through the overlay before actively looking for a local minimum.

3.2 The basic protocol

For clarity of the exposition, we first present a slightly simplified version of the LMS protocol. In this version, placing and locating items are essentially identical. We will then refine this to account for the differences between the two operations and to add optimizations.

To select a random local minimum, a node generates a *probe* message, which has the general form ⟨`probe`, *initiator, key, walk_length, path*⟩. A probe moves through the network with a *random walk* followed by a *deterministic walk*. The *walk_length* parameter is initialized to some positive value. A node that receives the probe (including the initiator) first examines this parameter. If it is greater than zero, the probe is in the random walk phase, and the node forwards it to a randomly selected neighbor[3] and decrements the value. If *walk_length* is zero, the probe is forwarded according to the deterministic walk, which is described below. In either case, a node appends itself to *path* before forwarding the probe so that a local minimum can direct its response back to the initiator.[4]

Deterministic walk Deterministic forwarding is done using a greedy algorithm. A node v receiving a probe computes the distance between *key* and every node in its neighborhood including itself. We define the distance between a key x and a node w as $d(x,w) = \min\{x - w \bmod 2^\lambda, w - x \bmod 2^\lambda\}$.

Let v' be the node that minimizes $d(key, \cdot)$ over the neighborhood of v. If $v = v'$, then v is the local minimum; it then stores or returns a replica, depending on the type of probe. Otherwise, v determines the next-hop node towards v' (from its local knowledge of the topology) and forwards the probe to this node. Recall that we assume v cannot communicate directly with v' if it is more than one hop away.

[1] This assumption can be weakened and communications made possible over larger distances in the overlay. This does not substantially alter the protocol, other than to allow optimizations that reduce the number of messages exchanged.

[2] For instance, the SHA-1 hash of a node's public key.

[3] The distribution for this selection is uniform in our case, but nonuniform choice is also possible. For instance, if the topology is dictated by trust relations, more highly trusted nodes might be given greater weight in selection. Another possibility, introduced in [6], is to weight the distribution according to resource availability.

[4] If messages can be exchanged by nodes that are not direct neighbors, then *path* is not needed, since the local minimum can communicate directly with the initiator.

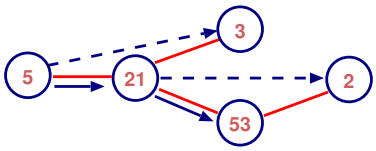


Figure 2: An illustration of "unexpected" message forwarding. Node 5 sees 3 in its 2-hop neighborhood, and forwards the probe to 21, the next hop towards 3. Node 21 sees a better match in 2, and so forwards the probe to 53 rather than 3.

For a local minimum of a key x, v_x, we define the set of nodes that deterministically forward to this minimum as its *basin of attraction* (or *basin*). Note that while it is possible for a local minimum to "attract" nodes from outside of its neighborhood, it is also possible for the basin to be only the minimum itself. This is illustrated in Fig. 1. It is also possible that a message "intended" for a node v' will be redirected towards a better minimum v'' at an intermediate hop, as shown in Fig. 2.

3.3 Protocol details

Replica placement We now distinguish between probes for placing replicas and those for locating replicas. A `REPLICA-PLACE` probe is initiated by the owner of an item, and includes an additional field *item*. When a local minimum receives a `REPLICA-PLACE` probe, it checks if it already holds a replica, and if not whether it has the resources to store one. If the minimum is able to store a replica of the item, it informs the initiator.

When a local minimum cannot store a new replica, it performs *duplication avoidance*: doubling the initial random walk length and restarting the probe's random walk. Because the probe is starting from a new location, it is less likely to again reach a duplicate minimum than if it were started with the same random walk length from the initiator.

We must limit the number of times that duplication avoidance is invoked, because the owner of an item might attempt to place more replicas than there are local minima for the item's key. This leads to `REPLICA-PLACE` probes that circulate through the network indefinitely, with progressively longer random walks. Consequently, we include a parameter initialized with the maximum number of failures permitted. Duplication avoidance decrements this parameter, discarding the probe when it becomes negative. The final form of a `REPLICA-PLACE` probe is then ⟨`REPLICA-PLACE`, *initiator, key, walk_length, item, initial_walk_length, failure_count, path*⟩.

Replica lookup A `SEARCH-PROBE` performs a lookup operation on an item key. A local minimum that receives a `SEARCH-PROBE` checks whether it holds a replica of the item and returns either the replica or notice of failure to the initiator. The form of a `SEARCH-PROBE` is simply ⟨`SEARCH-PROBE`, *initiator, key, walk_length, path*⟩.

Note that LMS is a probabilistic algorithm. The probability of a single `SEARCH-PROBE` locating a replica is fairly low, so a node will have to initiate a number of probes when looking for an item. This can be done in serial or in parallel. The former increases the expected time until the node receives a successful response, while the latter increases the load on the system.

3.4 Variant: Bloom filters

In this section, we describe a variant of the protocol, called the *Bloom filter variant*, which improves the efficiency of item lookup at the expense of additional communications overhead.

Each node periodically constructs a Bloom filter [5] of the keys for which it holds a replica and provides a copy of this Bloom filter to all the nodes in its h_B-neighborhood, where h_B is a global parameter. It is possible for h_B to be different than h, the depth of the neighborhood for the definition of local minimum (see Section 3).

Use of the filters within LMS is relatively straightforward: upon receiving a `SEARCH-PROBE` message, a node v can check all of the Bloom filters it holds for each of its h_B-hop neighbors. If there is a match of the key being searched for in the Bloom filter of some neighbor u, v can short-circuit the usual search protocol and forward the `SEARCH-PROBE` message directly to u.

Note that it is possible to use an *attenuated* Bloom filter [26] to combine incoming filters and transmit only a single filter for each distance. Thus, from each neighbor v a node receives a single filter for the items replicated at v, a single filter for all of v's one hop neighbors, a single filter for v's two hop neighbors, and so on. It is possible to further reduce the filter overhead by using arithmetic encoding to compress the filters (as described in [21]).

In general, it is not necessary to keep (attenuated) Bloom filters for the entire neighborhood for which a node keeps ID information (i.e. in general one can choose h_B to be smaller than h). Often it is sufficient to choose h_B to be 1 or 2 to get significant benefit without incurring the (computation and bandwidth) overhead of constructing and disseminating large filters from multiple hops away.

In the special case that the Bloom filter and identifier neighborhood are identical (i.e., $h_B = h$), the search phase of the protocol can dispense with the deterministic walk (Section 3.2). More specifically, if $h_B = h$, the Bloom filter variant of LMS mandates that search probes only perform a random walk. The search probe fails if *walk_length* reaches 0 without finding a Bloom filter match at any node on the path. We present an analysis of this special case, along with a general analysis of Bloom filters within LMS, in Appendix F.

3.5 The adaptive protocol

Having only local information, a node has no *a priori* way to know how many replicas of its items it must place in order for them to be easily found by other nodes. Increasing the number of replicas reduces the average number of search probes needed to find a replica, and vice-versa. Search performance thus provides a feedback mechanism to determine if an item has been replicated sufficiently.

The *adaptive protocol* is the component of LMS that determines and dynamically adjusts the number of replicas to be placed. It is run periodically for each item by its owner and ensures that a sufficient number r of replicas are available at any time. Let s be the average number of probes needed to find the item at a given time. The adaptive protocol ensures that $s \approx f(r)$, where f is an arbitrary function, chosen by the owner.[5] For simplicity of the exposition we present the protocol for the case $f(r) = r$ (see Section 4). The overhead of this protocol is minimal, aside from any additional replica placements, as it leverages existing actions.

A search from a random node in the system takes on average s probes to find a replica of an item I owned by node v. Node v periodically learns an estimate of s, as explained below; from the current number of reachable replicas r, it computes the number of replicas r' that it adaptively determines are needed as $r' = \lceil \alpha r + (1 - \alpha)s \rceil$. Here α is a hysteresis parameter between 0 and 1 (we use 0.9) that controls how sensitive the algorithm is to fluctuations in s. We define $\delta = r' - r$ as the needed change in number of replicas, and a non-zero value results in either placing or deleting replicas.

Since searches are a normal and frequent part of the system's operation, it is a simple matter for a node v_i to keep track of the number of probes s_i it needed to send before finding a replica of I.[6] If s_i is included in a `SEARCH-PROBE` message, the replica holder can store it and forward a batch of probe results to v periodically. These results provide a representative sample from which v can estimate the average s. Note that nodes on the threshold of being able to find replicas of I will tend to drive r up, allowing new nodes to find replicas and improving the sampling over the network.

[5] For example, this could be used to take the popularity of an item into account by storing many more replicas so that most searches require very few probes.

[6] For an unpopular item, the owner of an item might select random nodes and request that they perform searches for the item to make up for the lack of search feedback.

4 Analysis of LMS

We now give a rigorous analysis of the protocol; we make no assumptions about the topology layer, other than the necessary condition that the topology, which is a graph G, is connected. The four main results derived are: (a) tight bounds on the probability that a search in LMS fails; (b) bounds on the expected walk-length (i.e. the number of nodes visited by a search probe, Section 3.2) conditional on successful search; (c) the asymptotic performance of LMS; and (d) robustness under failures.

In Appendix F, we briefly analyze a variant of the protocol that includes Bloom filters.

4.1 Mixing time and eigenvalue gap

We first define the *mixing time* of a graph and the *eigenvalue gap* of the random walk. We will use these concepts to analyze the performance of LMS.

Let G be a connected directed graph and let n denote the number of nodes in G. Let $\mathcal{RW}(u,t)$ be the distribution of the vertex visited by a random walk on graph G after t steps, starting at vertex u. Only for the purpose of this analysis, we define the random walk in a way that is slightly different than what we did in the protocol description (Section 3.2): at each step, the walk remains at the current vertex with probability $1/2$ and otherwise moves to a uniformly random neighbor. Let P be the probability transition matrix of the random walk. Namely, $P[i,j]$ $(1 \leq i,j \leq n)$ is the probability that the random walk will move from node i to node j, at any step: for all edges (i,j) of G, it is $P[i,j] = 1/(2d_i)$, where d_i is the degree of node i; for all nodes i, $P[i,i] = 1/2$; $P[i,j] = 0$ otherwise. We denote by $1 = \mu_1 > \mu_2 \geq \cdots \geq \mu_n \geq 0$ the eigenvalues of P. We define the *eigenvalue gap* of the random walk as $g(G) = 1 - \mu_2$.

It is a known fact in graph theory that $\mathcal{RW}(u,t)$ approaches a unique *stationary distribution* $\mathcal{SRW}$, *no matter what the starting vertex u is.* $\mathcal{SRW}$ places probability $d(v)/(2m)$ on each node v, where $d(v)$ is the degree —number of neighbors— of v, and m is the total number of edges in G. We want to formalize this notion.

We can define the ϵ-mixing time of G, denoted by $T(G,\epsilon)$, as the smallest integer t such that for any vertex u it is:

$$\mathsf{SD}(\mathcal{RW}(\mathsf{u},\mathsf{t}),\mathcal{SRW}) \leq \epsilon/2$$

where $\mathsf{SD}(\mathsf{D},\mathsf{D}') = \frac{1}{2}\sum_{\mathsf{i}} |\Pr_{\mathsf{D}}[\mathsf{i}] - \Pr_{\mathsf{D}'}[\mathsf{i}]|$ denotes the statistical difference between the distributions D and D'.

It is known how to express the ϵ-mixing time as a function of the *eigenvalue gap*. To see this, we can apply a technique analogous to the proof of Theorem 6.21 in [22]: the statistical difference between $\mathcal{RW}(u,t)$ and $\mathcal{SRW}$ is at most $\sqrt{n}(\mu_2)^t$. From this, we obtain the following expression for the ϵ-mixing time:

$$T(G,\epsilon) \leq \frac{1}{g}\left[\frac{1}{2}\ln n + \ln(1/\epsilon)\right]. \tag{1}$$

For most natural models of randomly chosen/evolving graphs (such as random graphs and random regular graphs), it is known that the eigenvalue gap is lower bounded by a positive constant, independent of G and n (Appendix D). For such classes of graphs, $T(G,\epsilon) = O(\log n + \log(1/\epsilon))$ suffices. There exists, however, pathological graphs in which $T(G,\epsilon) = \Omega(n)$. This happens, for example, if G is a path.

In LMS, we take our parameter `INITIAL-WALK-LENGTH` to be at least $T(G,\epsilon)$, for a sufficiently small ϵ; in our simulations, choosing `INITIAL-WALK-LENGTH` to be 3 and doubling its value each time a duplicate local minimum is found is sufficient to obtain very good results.

4.2 Probability of successful search

Consider an item x. Let r denote the number of replicas of x that the owner of x places, and let s be the number of search probes that another node conducts to search for x. We derive an upper bound on the probability that the search fails. If G is regular and "symmetric" (e.g., if it is a random regular graph), we expect the r replicas of a key x to be placed uniformly at random at the k local minima for x; similarly for

the s locations at which the s searches end. In such a case, the probability of "failure" (unsuccessful search) is essentially of the form $(1 - \frac{r}{k})^s \leq \exp(-\frac{rs}{k})$, where $\exp(a)$ denotes e^a. However, we desire protocols that work for time-varying and arbitrary topologies G where the distribution may be quite non-uniform. Our first main result is that the failure probability is always bounded by a function of the form $\exp(-\Omega(rs/k))$, in the case of interest where r and s are bounded by $O(k)$. (Indeed, if r or s is $\Omega(k)$, we get a trivial linear-time algorithm.)

Theorem 4.1. *Let G be any connected undirected graph, on which we run LMS. Let u and v be two arbitrary nodes in G. Let x be an arbitrary identifier. Let $K(x)$ be the random variable representing the number of local minima of G with respect to a random id assignment to the nodes of G. Suppose node u publish an item with id x, using r replicas, and suppose that, later, node v look up an item with id x using s search probes. Assume that all random walks are of length at least $T(G, \epsilon)$. Let p_f be the probability that v fails to find the item. Then, conditioned on the event $K(x) = k$, the following bounds on the failure probability hold:*
(i) $p_f \leq s\epsilon + \exp(-(s^2/k) \cdot (1 - s/k))$, if $s = r$;
(ii) $p_f \leq s\epsilon + \exp(-\Omega(s \cdot \min\{r/k, 1\}))$, if $s \leq r$; and
(iii) the bound of (ii) with s and r interchanged, if $r < s$.

Note: The bound of (i) becomes trivial if $s \geq k$; in such cases where $s = r = \Omega(k)$, we can use (ii) to get a bound of $s\epsilon + \exp(-\Omega(s))$.

In practice, we limit ourselves to the case $r, s \leq k$ and we choose the parameters of LMS to obtain a small constant probability of failure. Therefore, Theorem 4.1 tells us that we need random walks of length $T(G, O(1/s)) = O((\log n)/g)$; as mentioned above, this is $O(\log n)$ for many realistic networks. We also need to choose r and s such that $rs = \Omega(k)$. We expand on this issue below.

We now present the main ideas of the proof of Theorem 4.1. A full proof appears in Appendix A.

For simplicity, let's focus on the case that ϵ is so small that can be neglected. Since all the random walks are chosen to be of length $t \geq T(G, \epsilon)$, we can pretend that the distributions $\mathcal{RW}(u, t)$ of the random walk during publishing and $\mathcal{RW}(v, t)$ of the random walk during lookup are both equal to the stationary distribution $\mathcal{SRW}$.

Our goal is to show that even when r and s are only moderately large, a search for x will succeed with high probability, no matter what G is. We proceed as follows. Let D be the probability distribution of choosing a vertex w (which is a local minimum for x) as follows: choose a vertex v using distribution $\mathcal{SRW}$, and then deterministically go to a local minimum w starting at v, as described in Section 3. Then, replica-placement is as follows: choose a multiset R of r local minima by sampling r times independently from D, remove duplicates from R, and place the replicas at the locations in R. Search for x is as follows: choose a multiset S of s local minima by sampling s times independently from D, and search is successful iff S intersects R.

We arrive at the following question: let $K = \{1, \ldots, k\}$ be the set of k local minima for x, and suppose we choose multisets R and S as in the previous paragraph. When can we show that $\Pr[R \cap S \neq \emptyset]$ is high? The heuristic argument a few paragraphs above shows that if D is the uniform distribution on K, then, the probability of unsuccessful search is of the form $\exp(-\frac{rs}{k})$; the following theorem shows that this upper bound indeed holds for any distribution D if $r, s \leq k$. This completes the proof sketch of Theorem 4.1.

Theorem 4.2. *Let r, s, k be three integers. Let D be an arbitrary distribution over the set $K = \{1, \ldots, k\}$. Let R be a set obtained by taking r independent samples from D and let S be a set obtained by taking s additional independent samples from D. Then:*
(i) $\Pr[R \cap S = \emptyset] \leq \exp(-(s^2/k) \cdot (1 - s/k))$, if $s = r$;
(ii) $\Pr[R \cap S = \emptyset] \leq \exp(-\Omega(s \cdot \min\{r/k, 1\}))$, if $s \leq r$; and
(iii) the bound of (ii) with s and r interchanged, if $r < s$.

The complete proof of this theorem appears in Appendix B. Here we simply make an observation. It seems reasonable to conjecture that D being the uniform distribution is the worst case (that is the case in which $\Pr[R \cap S = \emptyset]$ is highest), since otherwise elements of K that have high probability could appear in both R and S with increased likelihood. This intuition turns out be false: consider the case where $k = 2$, $s \ll r$, and D places probabilities p and $1 - p$ on the two elements of K. The failure probability here is

	Search cost	State
general case	$O(\frac{n}{rd_h}(\log n)/g)$	$O(d_h)$
$r = d_h = \Theta(n^{1/3})$	$O(n^{1/3}(\log n)/g)$	$O(n^{1/3})$

Table 1: Asymptotic performance of LMS. n is the number of nodes, d_h is the minimum size of a h-hop neighborhood and r is the number of replicas.

$p^r(1-p)^s + p^s(1-p)^r$. If D is the uniform distribution (i.e., $p = 1/2$), then the failure probability is $2^{1-(r+s)}$; however, the failure probability can be made much larger — about $(s/(r+s))^s \cdot \exp(-rs/(r+s))$ — by setting $p = s/(r+s)$. Similar counterexamples can be constructed for all k. In general, the case where r and s are quite different, needs care. We refer the reader to Appendix B for the details.

4.3 Expected walk-length

A second quantity of interest is the number N of nodes visited by any single search probe. This, multiplied by the number s of probes, yields the total cost of a search query (similarly, we get a multiplier of r in the replica-placement). We can write $N = T + l$, where T is the length of the random walk and l is the length of the deterministic walk. As we discussed above, $T = O((\log n)/g)$, where g is constant for practical networks. As far as l is concerned, we present the following result which shows that $l = O(\log n)$ with high probability:

Theorem 4.3. *For any graph G of n nodes and for any positive constant c, the number l of steps of any deterministic walk to a local minimum is at most $(c+1)\log n$ with high probability (at least $1 - 2/n^c$).*

The proof of this is deferred to Appendix C.

4.4 Asymptotic performance of LMS

Part (ii) of Theorem 4.1 allows us to estimate the performance of LMS. In particular, for $r, s \leq k$, the probability that a search fails is of the form $\exp(-\Omega(rs/k))$. Therefore, for any given k, we can choose $rs = \Theta(k)$ to make the failure probability an arbitrary small constant. For example, if $r = s = \lceil 2\sqrt{k} \rceil$, the probability of unsuccessful search is essentially at most $e^{-4} \sim 0.018$. (In fact, this upper bound only holds for relatively regular graphs; as G gets more irregular, the failure probability becomes smaller.)

Note that the number of local minima is, *in expectation*, $k = O(n/d_h)$, where d_h is the minimum h-hop neighborhood size in the graph. This means that, in any class of graphs where $d_h = d_h(n)$, if LMS places $r = r(n)$ replicas, then a search requires $s = O(\frac{n}{rd_h})$ probes. This translates into a search cost of $O(\frac{n}{rd_h}(\log n)/g)$ by virtue of Theorem 4.3 and Eq. (1). As far as the *routing state* kept by a node, this is proportional to the size of its h-hop neighborhood. Since this can be much larger than d_h, we assume that a node has a cap $\Delta = O(d_h)$ on the number of nodes in its neighborhood it is willing to keep state for.

Table 1 summarizes the asymptotic performance of the protocol. The table also shows, for concreteness, the interesting special case, where $d_h = \Omega(n^{\frac{1}{3}})$ (i.e., the given topology has h-hop neighborhoods that are not too small) and where we choose r such that r and s are of the same order of magnitude.

4.5 Fault-tolerance

We now investigate the performance of LMS when some of the replicas are lost. We consider two failure models: an *adversarial model* and a *random failure model*, that we now define. Let Y be the set of replicas found during a search for id x, assuming no faults thus far; i.e., if no faults have occurred thus far, the search has found $|Y|$ replicas of x.

The adversarial model is a very strong failure model that makes no assumptions about the timings of the failures: failures can even occur at the instant after the s searches have been completed. The adversary can arbitrarily delete up to t replicas of x, the adversary can wait until the searches have just terminated, and

Graph Type	Size	Avg. deg.	# Repl.	Visited LM	Visited LM+BF
power-law	10K	4.11	3	17.7	4.4
random	10K	4.11	22	131.1	21.8
Gnutella	61K	4.7	16	83.9	15.7
random	61K	4.7	45	282.8	43.8
random	100K	17	14	55.9	14.0
random	100K	12	19	87.1	19.0
random	100K	7	34	185.4	34.0

Table 2: Lookup performance for different graphs. The average number of lookup probes is approximately equal to the number of replicas. Measured quantities are averaged over 10,000 trials for each of 60 generated graphs, excluding the Gnutella graph.

then delete up to t replicas from Y. Thus, the search succeeds iff $|Y| \geq t+1$. In the random failure model, instead, each replica survives independently with probability p, that is, each element of Y can get deleted with probability $1-p$.

In such failure models, we can show the following: if the expected value $E[|Y|]$ is at least as large as a necessary lower bound, our search will succeed with high probability. In particular, in the adversarial model, we require $E[|Y|] \geq t + c_1\sqrt{t}$ for a constant t, where c_1 is some constant, for a good probability of success; this is nearly best-possible in the sense that there is another constant c_2 such that if $E[|Y|] \leq t + c_2\sqrt{t}$, then the success-probability can be very low. Similarly, in the random-failure model, we have the near-optimal result that $E[|Y|] \geq c_3/p$ suffices. In the specific case of near-regular networks, these requirements say that: (i) in the adversarial model, an overhead of only $O(\sqrt{t})$ in the number of replicas and in search time suffices, as compared to the no-fault case; and (ii) in the random-faults model, an overhead of $O(1/\sqrt{p})$ suffices. All these results are obtained by using a negative-correlation result in the proof of Theorem 4.2 (Appendix B) with some large-deviation bounds [23].

5 Experimental results

In this section we present an experimental study of LMS using a simulation (described in Appendix E) both to demonstrate large-scale behavior of the algorithm and to compare its performance on various topologies and with different extensions to the base protocol. In addition, we present results from a complete implementation of the base protocol.

5.1 Lookup performance

The results of the first set of experiments are shown in Table 2. The Gnutella graph represents a crawl of the actual deployed system [27, 28], while the others are generated. One randomly generated item is replicated in each trial, and the number of replicas is determined by an iterative procedure so that the average number of lookup probes needed to find one replica is approximately equal to the number of replicas. The "Visited" column shows the cost of the search either using the basic LMS protocol (the "LM" column) or the Bloom filter variant with $h_B = 2$ (the "LM+BF" column, see Section 3.3). The costs listed are the total number of nodes visited by the lookup probes until a replica is located (including all unsuccessful probes, executed serially).

The 10K graphs were chosen for comparison with the results of [18]. The most successful protocol in [18], *check*, visits approximately 150 nodes (slightly more than our 130), but requires over 90 replicas in contrast to our 22. As the graphs grow larger, LMS continues to perform reasonably. On the Gnutella graph, in particular, LMS requires relatively few replicas and lookup probes; we expect this graph to be typical of many

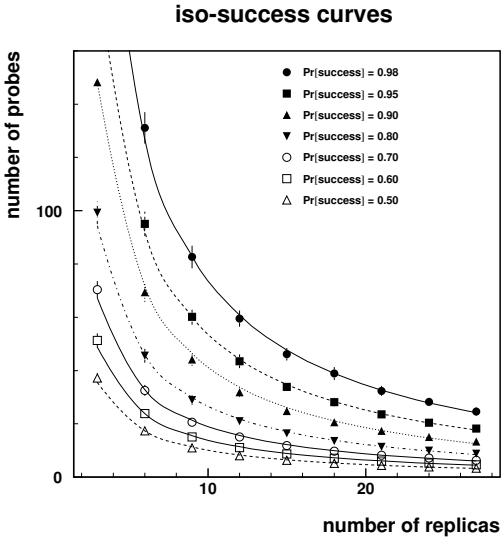


Figure 3: Number of probes as a function of the number of replicas. Random graph with 10^5 nodes, of average degree 17. The numbers labeling the curves are the particular success probabilities. The curves show fits to the functional form $f(r) = \frac{a}{r} + b$.

applications with unstructured topologies. Note that the Bloom-filter-based lookup performs considerably better than the standard LMS lookup, but this comes at the cost of maintaining and propagating Bloom filters.

In this experiment, we assume homogeneous nodes we do not take into account the effects of node congestion. Further our model does not allow a node to add neighbors. Therefore we do not discuss a comparison with Gia here (see Section 2).

Random graphs show the worst-case performance of LMS. In order to demonstrate the lower-bound behavior of the protocol, we restrict our experiments to these graphs hereafter.

5.2 Number of replicas vs. lookup overhead

In this section, we further consider details of LMS performance. We introduce the notion of an *iso-success curve*. An *iso-success* curve, for a given graph G and a given probability p, is the set of all pairs (r, s), such that if the owner of an item places r replicas and a node v uses *up to* s search probes to search the item, then the probability that v finds the item is p.

Iso-success curves precisely capture the tradeoff between number of replicas and lookup overhead. In Fig. 3, we present the iso-success curves for a single random graph of size 100K nodes (average degree 17). The figure shows the *worst-case* number of probes required for each number of replicas to achieve the given level of success probability (without using Bloom filters). For example, a little more than 20 probes were required with 10 replicas for 70% probability of finding an item. The curve was obtained by measuring the distribution of the number of probes for each number of replicas, using 10,000 trials.

For each success probability, the average number of search probes for each number of replicas was fit to the function $f(r) = \frac{a}{r} + b$, using the standard deviation of the numbers of probes as the uncertainty in their average. The fits demonstrate that $rs = Constant$ for a fixed probability is a very good approximation (b is small and negative in all fits).

A global fit to all of the data reveals that approximately half of the local minima dominate the network, greatly reducing the expected number of replicas needed for reliable lookup ($k/2$ rather than k in the results of Section 4).

Failure Probability	# Replicas	Average Visited	Analytic Bound
0	36	188	36
0.1	38	200	37.9
0.2	41	213	40.2
0.3	45	231	43.0
0.4	48	262	46.5
0.5	53	289	50.9

Table 3: Performance under random failures: random graph with 10^5 nodes, avg. degree of 7. The analytic bound predicts the number of replicas which in this case is equal to the number of probes.

Along with the tradeoff measure, the iso-success curves also provide us with a snapshot of the distribution of the expected number of probes required for lookups on a given graph. For example, with 15 replicas, 50% of the lookups require less than 8 probes (in the worst case), and only 5% of the lookups required more than 40 probes. This information is useful in implementation since it provides a strategy for fixing the number of probes that should be launched in parallel in order to reduce lookup latency.

5.3 Failure analysis

LMS is extremely robust, and in this section, we analyze its resilience under the failure model described in Section 4. Specifically, we consider random failures in which a given fraction of nodes in the networks fail. In these results, we consider how many replicas should be provisioned to handle specific failures (and validate our analysis). In these experiments, the adaptation is via static provisioning, i.e. the application/system designer places extra replicas. In later simulation (Section 5.4) and implementation results (Section 5.5), we consider how the system can adapt at run-time as it becomes aware of new failures.

In Table 3, we consider a random graph with 100,000 nodes (average degree 7). We consider that each replica, after being placed, can independently fail (the node discards the data) with the probability specified in the first column. In each case, we present an average over 10,000 runs of the number of replicas that needed to be placed *before the failures* such that the probability of a successful search *after the failures* is at least 99%. We present the specific placement that equalizes the number of search probes and the number of replicas.

If f is the failure probability, the number of replicas that equalizes the expected number of probes is $r(f) = \frac{r(0)}{\sqrt{1-f}}$. This analytic prediction is the last column of the table. It is clear that the analysis is extremely accurate in this case, and that LMS scales extremely well with failures. The most important point to note here is that the number of replicas only has to scale with the square root of the fraction of failures, which is an extremely positive outcome.

5.4 Performance under high churn

We now examine the behavior of LMS in a network with a high rate of nodes leaving and joining. The topology is a random graph of 10K nodes of average degree 7. A departing node is immediately replaced by a new node with the same number of neighbors; this maintains both the size of the graph and the degree distribution. Bloom filters are not used in this experiment. The neighborhood distance h is 2.

At the start of the simulation, a node u creates one replica of its item. Every minute (in simulation time) a random node searches for the item (over 99% of all searches succeed). The simulator also chooses q nodes at random (excluding u) to remove, where $q = 10000/T$ for an average node lifetime of T minutes. Every three minutes, the item's owner u selects a node by initiating a random walk (of length 6) and has that node perform a search for the item. Based on the result of the search, u applies the adaptive protocol (Section 3.5); placing or removing replicas as needed. The number of replicas r needed is based on a *cost*

Cost Ratio	Node Lifetime	Search Cost	Adaptive Overhead	Average Replicas
2	15	44.2	69.8	16.4
	30	35.9	63.4	20.3
	60	37.2	53.2	18.6
5	15	29.6	116.6	27.4
	30	35.7	70.7	23.8
	60	32	62.1	25.8

Table 4: Cost of searching, publishing and maintaining the overlay in basic LMS, in presence of high node churn. Lifetimes are measured in minutes.

ratio, an input parameter defined as r/s, where s is the expected number of probes to find the item[7]. The adaptive protocol then attempts to equalize the replication cost and the total cost of all (expected) searches.

The results of the experiment are shown in Table 4, which averages results over seven initial random graphs. The search cost is collected from the one-minute lookups. The adaptive overhead includes the cost of the search initiated by u and the cost of placing additional replicas. The search cost and number of replicas are fairly stable over all scenarios. The adaptive overhead is also relatively stable, with the exception of 15-minute lifetimes for a cost ratio of 5. Note that we do not include the overhead of maintaining the network, as it is external to the LMS protocol.

5.5 Implementation

We have implemented the complete LMS protocol (without Bloom filters). Multiple nodes are run on a single host, enabling us to test fairly large networks. The topology is maintained by a separate topology server that creates edges between nodes at random while enforcing a minimum degree for each node. All nodes are fail-stop; as nodes leave the network their neighbors obtain replacements as needed to maintain the minimum degree. The available bandwidth for each host and the capacity of the topology server define the limits of the network size that we were able to construct. The deployment was limited to locally available hosts.

Two experiments were performed, one with 512 nodes and one with 1024, both focusing on the behavior of the adaptive protocol, and both with a minimum node degree of 3 (average 3.9 ± 1.1) and $h = 2$. For the smaller network, we investigate a system in which searches and adaptation are frequent. The purpose of this experiment is to demonstrate that our protocols are functional in a real deployment under high (for the limited resources available) load. The behavior of the adaptive protocol is shown in Figure 4. Each node replicates one item[8] and performs a search for a random item every 10 seconds (staggered by random offsets at start-up), so that a new search is started roughly every 20 milliseconds. The adaptive protocol is run every 20 seconds (again staggered), and is based on the results of these searches. The average number of replicas tends to converge on a value dependent on the specific topology. The standard-deviation error band around the average number of replicas is dominated by the variation in the number of search probes, which we expect to be close to 1, as observed.

LMS probes increase in size as they propagate, but almost all are 400 bytes or smaller (excluding item size), and all are less than 1000 bytes. A successful lookup sends an "adaptive hint" back to the item's owner, the average size of which is 28.5 ± 0.5 bytes; most of this is the key identifier and can be batched by the replica holder or otherwise amortized. Replicas are soft state, so the owner of an item must send a refresh message (66.9 ± 0.8 bytes on average) to replica holders, which is done every 20 seconds in this

[7] In the notation of Section 3.5, this means running the adaptive protocol with $f(r) = r/(\text{cost ratio})$.

[8] Note that the number of items per node is not important when considering network load. What matters is the number of probes circulating through the network, or essentially the number of searches initiated per second. Our experiments are fundamentally limited by having eleven I/O-intensive processes on a single host, forcing us to restrict the rate at which searches are initiated.

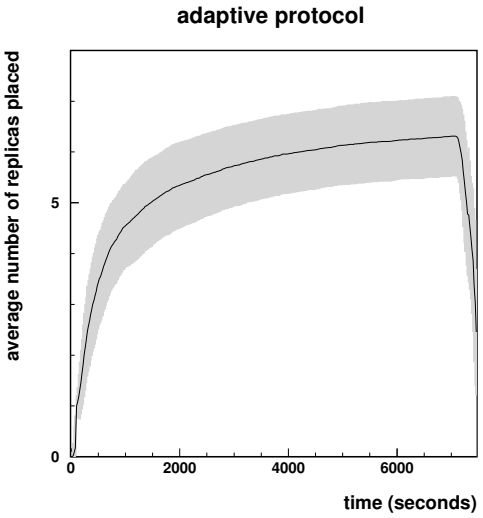


Figure 4: Effect of the adaptive protocol. The x axis shows elapsed time, with 20s between subsequent runs of the adaptive protocol. The y axis shows the average number of replicas placed for all items in the system. The shaded region indicates the one standard deviation region. The steep drop-off at the right is due to the system shutting down.

experiment. Replicas that are not refreshed are garbage-collected by their holder. Neighborhood propagation consumes considerable bandwidth, but we have constructed a much more efficient protocol that only propagates changes.

The experiment with 1024 nodes was used to verify that the adaptive protocol can cope with the sudden loss of half the network. Only one item is replicated, and its owner initiates periodic searches for it. Halfway through the experiment 512 of the nodes depart. We do not present specific results of this, as providing no special insight into the system, but mention that it performed very well under extremely adverse conditions.

6 Conclusions

In this paper, we designed LMS, an efficient unstructured P2P protocol that provides a DHT-like publish and lookup functionality. Unlike a DHT [31, 29, 25], LMS is designed for a model in which the topology (i.e. the graph where a vertex denotes a peer and an edge denotes that two peers know of each other and can communicate directly) is determined by an external entity, therefore peers are not allowed to choose their neighbors. We demonstrated through analysis and simulation that LMS is an efficient protocol with stronger performance guarantees than other unstructured protocols that work in the same model. We also presented a prototype implementation, which shows the practicality of the design.

LMS is of practical interest for distributed applications relying on trust relations between nodes. These trust relations define a graph which, when traversed along its links, provides a known level of assurance for operations. Specifically, we developed a decentralized public-key infrastructure based on a web-of-trust. Traversing links in the trust graph implicitly generates certificate chains assuring a node searching for a public key of that key's correctness. More details on this application will be available in a later publication.

7 Acknowledgements

This material is based upon work supported in part by: NSF grant 0208005, ITR Award CNS-0426683, NSF grant CCR-0310499, NSF award ANI0092806 and DoD contract MDA90402C0428.

We thank the referees for their valuable comments.

Bobby Bhattacharjee and Aravind Srinivasan are also affiliated with the Institute for Advanced Computer Studies, University of Maryland.

References

[1] I. Abraham, C. Gavoille, D. Malkhi, N. Nisan, and M. Thorup. Compact name-independent routing with minimum stretch. In *SPAA '04: Proceedings of the sixteenth annual ACM symposium on Parallelism in algorithms and architectures*, pages 20–24, New York, NY, USA, 2004. ACM Press.

[2] I. Abraham and D. Malkhi. Probabilistic quorums for dynamic systems. In *International Symposium on Distributed Computing (DISC)*, pages 60–74, 2003.

[3] I. Abraham and D. Malkhi. Name independent routing for growth bounded networks. In *SPAA'05: Proceedings of the 17th annual ACM symposium on Parallelism in algorithms and architectures*, pages 49–55, New York, NY, USA, 2005. ACM Press.

[4] M. Arias, L. J. Cowen, K. A. Laing, R. Rajaraman, and O. Taka. Compact routing with name independence. In *SPAA '03: Proceedings of the fifteenth annual ACM symposium on Parallel algorithms and architectures*, pages 184–192, New York, NY, USA, 2003. ACM Press.

[5] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, July 1970.

[6] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker. Making Gnutella-like P2P systems scalable. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 407–418. ACM Press, 2003.

[7] J. Chu, K. Labonte, and B. N. Levine. Availability and locality measurements of peer-to-peer file systems. In *Proc. ITCom: Scalability and Traffic Control in IP Networks II Conferences*, volume 4868, July 2002.

[8] I. Clarke, S. G. Miller, T. W. Hong, O. S. Sandberg, and B. Wiley. Protecting free expression online with Freenet. *IEEE Internet Computing*, 6(1), Jan./Feb. 2002.

[9] D. Dubhashi and D. Ranjan. Balls and bins: a study in negative dependence. *Random Struct. Algorithms*, 13(2):99–124, 1998.

[10] J. Friedman. A proof of alon's second eigenvalue conjecture. In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 720–724, New York, NY, USA, 2003. ACM Press.

[11] A. M. Frieze. Edge-disjoint paths in expander graphs. In *SODA '00: Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, pages 717–725, Philadelphia, PA, USA, 2000. Society for Industrial and Applied Mathematics.

[12] P. Ganesan, Q. Sun, and H. Garcia-Molina. Yappers: A peer-to-peer lookup service over arbitrary topology. In *22nd Annual Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM)*, 2003.

[13] D. Gillman. A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220, 1998.

[14] C. Gkantsidis, M. Mihail, and A. Saberi. Random walks in peer-to-peer networks. In *IEEE Infocom*, 2004.

[15] `http://www.gnutella.com`.

[16] S. Jiang, L. Guo, and X. Zhang. LightFlood: an efficient flooding scheme for file search in unstructured peer-to-peer systems. In *International Conference on Parallel Processing*, 2003.

[17] N. Kahale. Large deviation bounds for Markov chains. Technical Report 94-39, DIMACS, 1994.

[18] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker. Search and replication in unstructured peer-to-peer networks. In *Proceedings of the 16th international conference on Supercomputing*, pages 84–95. ACM Press, 2002.

[19] D. Malkhi, M. Naor, and D. Ratajczak. Viceroy: a scalable and dynamic emulation of the butterfly. In *Proceedings of the twenty-first annual symposium on Principles of distributed computing*, pages 183–192. ACM Press, 2002.

[20] A. Medina, I. Matta, and J. Byers. On the origin of power laws in Internet topologies. *ACM SIGCOMM Computer Communication Review*, 30(2):18–28, 2000.

[21] M. Mitzenmacher. Compressed Bloom filters. In *20th Annual ACM Symposium on Principles of Distributed Computing (PODC '01)*, pages 144–150, Newport, RI USA, August 26–29 2001. ACM Press.

[22] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1997.

[23] A. Panconesi and A. Srinivasan. Randomized distributed edge coloring via an extension of the Chernoff-Hoeffding bounds. *SIAM J. Comput.*, 26(2):350–368, 1997.

[24] V. Ramasubramanian and E. G. Sirer. Beehive: Exploiting power law query distributions for O(1) lookup performance in peer to peer overlays. In *Proceedings of NSDI*, 2004.

[25] S. Ratnasamy, P. Francis, M. Handley, R. M. Karp, and S. Shenker. A scalable content-addressable network. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '01)*, pages 161–172, San Diego, CA USA, August 27–31 2001. ACM.

[26] S. C. Rhea and J. Kubiatowicz. Probabilistic location and routing. In *Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002), Proceedings*, volume 3, pages 1248–1257, New York, NY USA, June 23–27 2002. IEEE.

[27] `http://people.cs.uchicago.edu/~matei /GnutellaGraphs/`.

[28] M. Ripeanu and I. T. Foster. Mapping the Gnutella network: Macroscopic properties of large-scale peer-to-peer systems. In P. Druschel, M. F. Kaashoek, and A. I. T. Rowstron, editors, *Peer-to-Peer Systems, First International Workshop (IPTPS 2002)*, volume 2429 of *Lecture Notes in Computer Science*, pages 85–93, Cambridge, MA USA, March 7–8 2002. Springer.

[29] A. I. Rowstron and P. Druschel. Pastry: Scalable, distributed object location, and routing for large-scale peer-to-peer systems. In *Proceedings of the 18th IFIP/ACM International Conference on Distributed Systems Platforms (Middleware 2001)*, volume 2218 of *Lecture Notes in Computer Science*, pages 329–350, Heidelberg, Germany, 2001. Springer.

[30] S. Saroiu, K. P. Gummadi, and S. D. Gribble. Measuring and analyzing the characteristics of Napster and Gnutella hosts. *Multimedia Syst.*, 9(2):170–184, August 2003.

[31] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '01)*, pages 149–160, San Diego, CA USA, August 27–31 2001. ACM.

[32] J. Winick and S. Jamin. Inet-3.0: Internet topology generator. Technical Report UM-CSE-TR-456-02, University of Michigan, 2002.

[33] B. Zhao, K. Kubiatowicz, and A. 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, 2001.

A Proof of Theorem 4.1

We restate the theorem for convenience:

Let G be any connected undirected graph, on which we run LMS. Let u and v be two arbitrary nodes in G. Let x be an arbitrary identifier. Let $K(x)$ be the random variable representing the number of local minima of G with respect to a random id assignment to the nodes of G. Suppose node u publish an item with id x, using r replicas, and suppose that, later, node v look up an item with id x using s search probes. Assume that all random walks are of length at least $T(G, \epsilon)$. Let p_f be the probability that v fails to find the item. Then, conditioned on the event $K(x) = k$, the following bounds on the failure probability hold:
(i) $p_f \leq s\epsilon + \exp(-(s^2/k) \cdot (1 - s/k))$, if $s = r$;
(ii) $p_f \leq s\epsilon + \exp(-\Omega(s \cdot \min\{r/k, 1\}))$, if $s \leq r$; and
(iii) the bound of (ii) with s and r interchanged, if $r < s$.

Proof. For simplicity of exposition, we limit the proof to case (ii). The remaining two cases can be proved analogously.

Let t be the length of the random walks used. For any vertex y, let D_y be the probability distribution of choosing a vertex w (which is a local minimum for x) as follows: choose a vertex w' using distribution $\mathcal{RW}(y, t)$, and then deterministically go to a local minimum w starting at w', as described in Section 3. Then, replica-placement is as follows: choose a multiset R of r local minima by sampling r times independently from D_u, remove duplicates from R, and place the replicas at the locations in R. Search for x is as follows: choose a multiset S of s local minima by sampling s times independently from D_v, and search is successful iff S intersects R.

We introduce a hybrid experiment, where we choose a set S^* of s random samples according to the distribution D_u. Theorem 4.2 implies that:

$$\Pr[R \cap S^* = \emptyset] \leq B(r, s, k).$$

where $B(r, s, k) = \exp(-\Omega(s \cdot \min\{r/k, 1\}))$ is the bound given by Theorem 4.2 in case (ii).

We now have to relate $p_f = \Pr[R \cap S = \emptyset]$ with $\Pr[R \cap S^* = \emptyset]$. Since $t \geq T(G, \epsilon)$, this implies that the statistical difference between D_u and D_v is at most ϵ. Let D_u^s (resp. D_v^s) be the distribution consisting of s independent samples from D_u (resp. D_v). A straightforward hybrid argument shows that

$$\mathsf{SD}(\mathsf{D_u^s}, \mathsf{D_v^s}) \leq \mathsf{s}\epsilon$$

which implies that the statistical difference between the random variables S and S^* is also at most $s\epsilon$.

We now need the following claim:

Claim 1. *Let X_1 and X_2 are two random variable with statistical difference at most ϵ. Let $E(X, Y)$ be an event that depends only on the underlining random variable X, Y (i.e. E is a predicate of two variables). Then, for any distribution of random variable Y:*

$$|\Pr[E(X_1, Y)] - \Pr[E(X_2, Y)]| \leq \epsilon$$

Proof. Fix an arbitrary value y. We first show that the statement holds when conditioning on the event $Y = y$. The claim follows by unconditioning. Let:

$$A^+ = \{x : E(x,y) \text{ and } \Pr[X_1 = x] > \Pr[X_2 = x]\}$$
$$A^- = \{x : E(x,y) \text{ and } \Pr[X_1 = x] < \Pr[X_2 = x]\}$$

We get:

$$|\Pr[E(X_1,Y)] - \Pr[E(X_2,Y)]| = |\sum_{x\in A^+} \Pr[X_1 = x] + \sum_{x\in A^-} \Pr[X_1 = x] + \sum_{x\in A^+} \Pr[X_2 = x] + \sum_{x\in A^-} \Pr[X_2 = x]| =$$
$$\left|\sum_{x\in A^+} (\Pr[X_1 = x] - \Pr[X_2 = x]) - \sum_{x\in A^-} (\Pr[X_2 = x] - \Pr[X_1 = x])\right|.$$

This quantity is maximized when the first summation is maximum and the second is minimum. Specifically, for any fixed choice of the distributions of X_1 and X_2, the quantity is maximized by choosing E such that $A^+ = \{x : \Pr[X_1 = x] > \Pr[X_2 = x]\}$ and A^- is empty. Therefore:

$$|\Pr[E(X_1,Y)] - \Pr[E(X_2,Y)]| \leq \sum_{x:\Pr[X_1=x]>\Pr[X_2=x]} (\Pr[X_1 = x] - \Pr[X_2 = x])$$

and simple algebra shows that the right hand side of the last inequality is equal to $\mathsf{SD}(\mathsf{X}_1, \mathsf{X}_2)$, which is at most ϵ. This completes the proof. □

Using the Claim, we can write:

$$|\Pr[R \cap S = \emptyset] - \Pr[R \cap S^* = \emptyset]| \leq s\epsilon$$

from which case (ii) of the theorem follows. □

B Proof of Theorem 4.2

We restate the theorem for convenience:

Let r, s, k be three integers. Let D be an arbitrary distribution over the set $K = \{1, \ldots, k\}$. Let R be a set obtained by taking r independent samples from D and let S be a set obtained by taking s additional independent samples from D. Then:
(i) $\Pr[R \cap S = \emptyset] \leq \exp(-(s^2/k) \cdot (1 - s/k))$, *if* $s = r$;
(ii) $\Pr[R \cap S = \emptyset] \leq \exp(-\Omega(s \cdot \min\{r/k, 1\}))$, *if* $s \leq r$; *and*
(iii) the bound of (ii) with s and r interchanged, if $r < s$.

Proof. Before diving into the proof details, we first discuss a simple approach to the problem and we show why it does *not* work. One would be tempted to proceed as follows: denote the s samples from D that form the set S by the random variables $e_1, \ldots, e_s$ and try to express $\Pr[R \cap S =]$ as a function of $y = \Pr[e_i \notin R]$. We immediately note a stumbling block: given that one element of S did not lie in R, the conditional probability of this happening for another element of S can go up! Thus we have an undesirable positive correlation among these events: in particular, if y is the probability that an arbitrary single element of S does not lie in R, then the probability that R and S are disjoint is *at least as large as* y^s.

We take a different approach, wherein the correlations actually help us. Recall that $K = \{1, 2, \ldots, k\}$, and let X_i be the event that i lies in both R and S. As usual, $\overline{X_i}$ denotes the complement of X_i; letting p_i be the probability that distribution D places on i and setting $q_i = 1 - p_i$, we have $\Pr[\overline{X_i}] = (q_i^r + q_i^s - q_i^{r+s})$.

The probability that R and S are disjoint is $\Pr[\bigwedge_{i=1}^{k} \overline{X_i}]$. Our first key idea is that the events $\overline{X_i}$, $i = 1, 2, \ldots, k$, are negatively correlated! Intuitively, this seems clear: given that none of $X_1, X_2, \ldots, X_{i-1}$ held, this informally "leaves more slots free" for X_i to occur. We prove this by using a result by Dubhashi and Ranjan [9]. For $i = 1, \ldots, k$ and for $j = 1, \ldots, r + s$, let $B_{i,j}$ be the indicator of the event $\{\sigma_j = i\}$, where $\sigma_1, \ldots, \sigma_r$ are the samples from D that constitute R and $\sigma_{r+1}, \ldots, \sigma_{r+s}$ are the samples that constitute S. Proposition 12 of [9] tells us that the $B_{i,j}$ are *negatively associated*, as defined in Definition 3 of the same paper. For every i, we can write the event $\overline{X_i}$ as $f(B_{i,1}, \ldots, B_{i,r+s})$ and the event "$\bigwedge_{l=1}^{i-1} \overline{X_l}$" as $g((B_{l,j})_{l=1,\ldots,i-1;j=1,\ldots,r+s})$, where f and g are appropriate non-increasing functions. Applying the definition of negative association, we get that $\overline{X_i}$ as "$\bigwedge_{l=1}^{i-1} \overline{X_l}$" are negatively *correlated*, as needed. This negative correlation leads to the bound

$$\begin{aligned} \Pr[\bigwedge_{i=1}^{k} \overline{X_i}] &\leq \prod_{i=1}^{k} \Pr[\overline{X_i}] \qquad (2) \\ &= \prod_{i=1}^{k} (q_i^r + q_i^s - q_i^{r+s}). \qquad (3) \end{aligned}$$

Part (i) of the theorem ($r = s$) can now be proved by using the concavity of the function $z \mapsto \ln(f(z))$ over the domain $z \in (0, 1)$, where $f(z) = 2z^r - z^{2r}$. We can write:

$$\ln \Pr[\bigwedge_{i=1}^{k} \overline{X_i}] \leq \sum_{i=1}^{k} \ln(f(q_i)). \qquad (4)$$

Note that $\sum_i q_i = k - \sum_i p_i = k - 1$. Now, by considering the second derivative, it can be shown that in the domain $z \in [0, 1]$, the function $z \mapsto \ln(f(z))$ is *concave*. Thus, subject to the constraint $\sum_i q_i = k - 1$, the value $\sum_{i=1}^{n} \ln(f(q_i))$ is maximized when all the q_i are equal (i.e., equal to $1 - 1/k$). This, in combination with (4) and some further calculation, leads to the bound of part (i) of the theorem.

We now consider part (ii) of the theorem, where $s \leq r$. As shown in the example with $k = 2$ in Section 4, this needs more care. In particular, the analog of the function f above, can have the property that $\ln f$ is neither totally convex nor totally concave in the domain $z \in (0, 1)$. The basic ideas here are as follows. Recall the probabilities p_i from above. Partition K into three sets:

$$\begin{aligned} C_1 &= \{i : p_i \leq 1/r\}; \\ C_2 &= \{i : 1/r < p_i \leq 1/s\}, \text{ and} \\ C_3 &= \{i : p_i > 1/s\}. \end{aligned}$$

The product in (3) now splits into three products, which we will analyze separately and then combine. For $j = 1, 2, 3$, let $v_j = \sum_{i \in C_j} p_i$ and note that the sum $v_1 + v_2 + v_3 = 1$.

Let us first consider any $i \in C_1$. Since $0 \leq p_i \leq 1/r$, it is easy to show here that

$$q_i^r + q_i^s - q_i^{r+s} = 1 - \Theta(rsp_i^2).$$

(To be accurate, we bound the quantity q_i^t from below and above for any $t = r, s, r + s$:

$$\begin{aligned} q_i^t &= (1 - p_i)^t \geq 1 - tp_i + \tfrac{t(t-1)}{2} p_i^2 - \tfrac{t(t-1)(t-2)}{6} p_i^3 \\ q_i^t &= (1 - p_i)^t \leq 1 - tp_i + \tfrac{t(t-1)}{2} p_i^2. \end{aligned}$$

For $t = r + s$, using the fact that $tp_i \leq 2$:

$$\begin{aligned} q_i^t &\geq 1 - tp_i + \tfrac{t(t-1)}{2} p_i^2 - \tfrac{(t-1)(t-2)}{3} p_i^2 \\ &\geq 1 - tp_i + \tfrac{(t-1)(t+4)}{6} p_i^2, \end{aligned}$$

from which:

$$\begin{aligned} q_i^r + q_i^s - q_i^{r+s} &\leq 1 - rp_i + \tfrac{1}{2}r^2p_i^2 - \tfrac{1}{2}rp_i^2 + 1 - sp_i + \tfrac{1}{2}s^2p_i^2 \\ &\quad -\tfrac{1}{2}sp_i^2 - 1 + (r+s)p_i - \tfrac{(r+s)^2+3(r+s)-4}{6}p_i^2 \\ &\leq 1 - \tfrac{2rs+2(r+s)-4}{6}p_i^2 \\ &\leq 1 - \tfrac{rs}{3}p_i^2. \end{aligned}$$

) Next, basic convexity arguments show that for any constant $\alpha > 0$, and any given values for $|C_1|$ and v_1, the quantity $\prod_{i \in C_1}(1 - \alpha rsp_i^2)$ is maximized (as a function of the variables p_i) when $p_i = v_1/|C_1|$ for each $i \in C_1$. Further simplification yields

$$\prod_{i \in C_1}(q_i^r + q_i^s - q_i^{r+s}) \leq \exp(-\Omega(rsv_1^2/|C_1|)). \tag{5}$$

(More precisely, the bound is $\exp(-\frac{rsv_1^2}{3|C_1|})$.)

Next consider C_2. In this case,

$$q_i^r + q_i^s - q_i^{r+s} = 1 - \Theta(sp_i).$$

(More accurately, we get:

$$\begin{aligned} q_i^r + q_i^s - q_i^{r+s} &\leq \\ &\leq (1-p_i)^s + (1-p_i)^{1/p_i}(1-(1-p_i)^s) \\ &\leq (1 - sp_i + s^2p_i^2/2) + sp_i/e \\ &\leq 1 - sp_i + sp_i/2 + sp_i/e \\ &= 1 - (\tfrac{1}{2} - \tfrac{1}{e})sp_i \leq 1 - 0.13sp_i. \end{aligned}$$

) A simple argument yields

$$\prod_{i \in C_2}(q_i^r + q_i^s - q_i^{r+s}) \leq \exp(-\Omega(s \cdot v_2)). \tag{6}$$

(More precisely, the bound is $\exp(-0.13sv_2)$.)

The set C_3 is where the function $q_i^r + q_i^s - q_i^{r+s}$ exhibits complex behavior. Fortunately, we can deal with it as follows. For any $i \in C_3$, $q_i^r + q_i^s - q_i^{r+s}$ is at most

$$q_i^r + q_i^s \leq 2 \cdot q_i^s \leq 2 \cdot \exp(-sp_i) \leq \exp(-\Omega(sp_i)),$$

(more precisely, $\exp(-0.3sp_i)$) where the final inequality follows from the fact that $p_i \geq 1/s$. Therefore,

$$\prod_{i \in C_3}(q_i^r + q_i^s - q_i^{r+s}) \leq \exp(-\Omega(s \cdot v_3)). \tag{7}$$

(More precisely, $\exp(-0.3v_3)$).

Putting (5), (6) and (7) together with (3), we get that for some constant $\gamma > 0$, $\Pr[\bigwedge_{i=1}^k \overline{X_i}]$ is upper-bounded by

$$\exp(-\gamma \cdot s \cdot (rv_1^2/|C_1| + v_2 + v_3)),$$

(where γ can be taken to be 0.13) which equals $\exp(-\gamma \cdot s \cdot (rv_1^2/|C_1| + 1 - v_1))$; this, in turn, is at most $\exp(-\gamma \cdot s \cdot (rv_1^2/k + 1 - v_1))$. Elementary calculus that determines the maximum of this last expression as a function of v_1 (where v_1 takes values in $[0,1]$), yields the bound of part (ii) of the theorem. (Specifically, there are two cases. If $r \leq k/2$, then the function is maximum for $v_1 = 1$, in which case the bound becomes $\exp(-\gamma rs/k)$. If $r > k/2$, the maximum is $v_1 = k/(2r)$, in which case the bound becomes $\exp(-\gamma s(1-k/(4r))$ which is at most $\exp(-(\gamma/2)s)$. In the special case $k/2 < r < k$ this can be further upper bounded by $\exp(-(\gamma/2)rs/k)$.)

Finally, part (iii) of the theorem follows by symmetry. □

C Proof of Theorem 4.3

We restate the theorem for convenience:

For any graph G of n nodes and for any positive constant c, the number l of steps of any deterministic walk to a local minimum is at most $(c+1)\log n$ with high probability (at least $1-2/n^c$).

Proof. Fix an arbitrary object x. We aim to show that the "deterministic walk to a local minimum for x" takes at most $O(\log n)$ steps in expectation, and also with high probability.

Recall that we currently hash to a very large universe of (say, 160-bit) IDs, and also recall our notion of distance between two hash values. Since the ID-space has size large enough (2^{160}), the situation is essentially equivalent to the following. Let C be a circle of unit circumference. We then hash each entity v (whether it is an object or a node) to a random point $h(v)$ on the circumference of C. The distance between two points p and q that lie on C, denoted $\Delta(p,q)$, is the length of the shorter of the two arcs that connect them; thus, $\Delta(p,q) \leq 1/2$. Note that this is the same notion of distance that we currently employ; thus, our notion of local minima etc. carries over here exactly. (That is, we always aim to find a neighbor that has the smallest distance $\Delta(\cdot,\cdot)$ to the value $h(x)$.) The utility of mapping on to the circle is that the analysis becomes smoother (e.g., via integrals).

Suppose we start at a node u_0, and are routing to a local minimum for x. We assume without loss of generality that the hash value $h(x)$ of x, is the lowest point P on C. Whenever we say "distance", we will mean the distance function $\Delta(\cdot,\cdot)$. We now introduce some random variables. Let $u_1, u_2, \ldots$ be the successive nodes visited; if u_i is the local minimum found, then u_{i+1}, u_{i+2} etc. equal u_i. Let $M_i = \Delta(h(u_i), P)$ denote the distance "between u_i and x"; note that the sequence $M_0, M_1, \ldots$ decreases until we hit a local minimum. Our key idea is to show that this sequence decreases fast enough. For $t \geq 0$, let A_t be the indicator random variable for the event that the walk has **not yet** stopped after t steps (i.e., after visiting $u_0, u_1, \ldots, u_t$).

Our key lemma is the following:

Lemma C.1. *For any $t \geq 1$ and any $z \in [0, 1/2]$, $\boldsymbol{E}[M_{t+1}A_t \mid M_tA_{t-1} = z] \leq z/2$.*

We will prove Lemma C.1 below; let us now see why the lemma yields the desired $O(\log n)$ bound. Note that $\mathbf{E}[M_1A_0] \leq \mathbf{E}[M_1] \leq 1/2$; thus, Lemma C.1 and an induction on t yield that $\mathbf{E}[M_tA_{t-1}] \leq 2^{-t}$. In particular, letting $T = (c+1)\log n$ where c is some suitable constant, we get

$$\mathbf{E}[M_TA_{T-1}] \leq n^{-(c+1)}. \tag{8}$$

Now, suppose $u_T = u$ and that $M_TA_{T-1} = z$. Node u has at most n unexplored neighbors; for each of these neighbors v, $\Pr[h(v) < z] = 2z$. (The factor of two comes from the fact that $h(v)$ can fall on either side of point P on the circle.) Thus, the probability that u is **not** a local minimum is at most $2nz$; more formally,

$$\forall z,\ \Pr[A_T = 1 \mid M_TA_{T-1} = z] \leq 2nz.$$

So,

$$\Pr[A_T = 1] \leq 2n \cdot \mathbf{E}[M_TA_{T-1}] \leq 2/n^c$$

by (8), which is negligible if, say, $c \geq 2$. Thus, the routing takes at most $O(\log n)$ steps in expectation, and also with high probability. □

We now prove Lemma C.1:

Proof. First of all, we can assume that $z \neq 0$; since $M_{t+1} \leq M_t$ and $A_t \leq A_{t-1}$, the lemma is trivial if $z = 0$. Therefore, we have $A_{t-1} = 1$ and $M_t = z$; i.e., the walk has not stopped after visiting node u_{t-1}, and also $\Delta(h(u_t), P) = z > 0$. Let Y be the random variable denoting the number of "unexplored" neighbors of u_t; i.e., the number of neighbors of u_t that do not belong to the set $\{u_0, u_1, \ldots, u_{t-1}\}$. If $Y = 0$, then u_t is a local minimum, and hence $M_{t+1}A_t = 0$. We will now prove that for all $d \geq 1$,

$$\mathbf{E}[M_{t+1}A_t \mid ((M_tA_{t-1} = z) \wedge (Y = d))] \leq z/2. \tag{9}$$

If we can do so, we will be done, since if the lemma holds conditional on all positive values of d, it also holds unconditionally. For notational simplicity, we will from now on refer to the l.h.s. of (9) as Φ.

Fix some $d \geq 1$; in all arguments below, we are conditioning on the event "$(M_t A_{t-1} = z) \wedge (Y = d)$". Let $v_1, v_2, \ldots, v_d$ denote the d unexplored neighbors of u_t. If $h(v_i) > h(u_t)$ for all i, then u_t is a local minimum, and hence $M_{t+1}A_t = 0$. Therefore, conditioning on the value $y = \min_i d(h(v_i), P) \leq z$, and also considering the d possible values of i that achieve this minimum, we get

$$\Phi = 2d \cdot \int_{y=0}^{z} y(1-2y)^{d-1} dy.$$

Again, the factor of two up-front comes from the fact that the "minimizing neighbor" v_i can fall on either side of point P on the circle. A simple computation yields

$$\Phi = \frac{1 - (1-2z)^d \cdot (1+2zd)}{2(d+1)}. \tag{10}$$

We need to show that the r.h.s. of (10) is at most $z/2$. Changing variables to $y := 2z$ and rearranging, we need to show that

$$f(y) \doteq (d+1)y/2 + (1-y)^d(1+yd) \geq 1,$$

where $0 \leq y \leq 1$ and $d \geq 1$ is an integer. This inequality is easily seen to hold if $d = 1$, so we may assume $d \geq 2$.

It can be verified that $f'(y)$ equals $(d+1)/2 + d(1-y)^d - d(1+yd)(1-y)^{d-1}$. Also, $f'(1/d)$ equals

$$(d+1) \cdot (1/2 - (1-1/d)^{d-1}), \tag{11}$$

and $f''(y)$ equals

$$d(1-y)^{d-2}(d+1) \cdot (dy - 1). \tag{12}$$

We see from (12) that f' has a unique minimum (in our domain $0 \leq y \leq 1$) at $y = 1/d$. Also, for integer $d \geq 2$, the function $(1-1/d)^{d-1}$ has value $1/2$ when $d = 2$, and decreases as d takes on higher integral values. Thus, (11) shows that $f'(1/d) \geq 0$. Since this minimum value is non-negative, it follows that $f'(y) \geq 0$ for $0 \leq y \leq 1$. So, since $f(0) = 1$, we get $f(y) \geq 1$ for all $y \in [0, 1]$, as required. □

D Mixing time in certain classes of graphs

In this appendix, we justify our claim that random and random regular graphs have constant eigenvalue gap, and therefore logarithmic mixing time, under reasonable assumptions.

D.1 Mixing time in random graphs

A random $G(n,p)$ graph has a constant eigenvalue gap, i.e. $g = \Omega(1)$, with high probability, if $p = \Omega((\ln n)/n)$. Although we believe this fact was previously known, we could not find a reference. Therefore, we provide a simple proof sketch.

Let Φ be the expansion factor of a graph. If $p \geq 4(\ln n)/n$, then the random graph $G(n,p)$ has an expansion factor of at least $.066np$, with high probability (at least $1 - 1/n$). We prove this, by showing that for all $t = 1, \ldots, n/2$, and for any set S of t nodes in the graph, the probability that S has too few outgoing edges (i.e. less than $(1-\delta)$ times the expected number of outgoing edges) is at most $\frac{1}{\binom{n}{t}}$. The proof is a standard application of the Chernoff bound.

Let P be the probability transition matrix of a random walk on a graph G, as defined in Section 4.1, and let $1 = \mu_1 > \mu_2 \geq \cdots \geq \mu_n \geq 0$ be its eigenvalues.

Frieze [11, Eq.(12,13)] gives the following bound on the eigenvalue gap $g = 1 - \mu_2$:

$$g \geq \frac{\Phi^2 \delta^2}{8\Delta^4},$$

where Φ is the expansion factor of G and δ and Δ are the minimum and maximum degree of G.

In case of $G(n,p)$, if $p \geq 8(\ln n)/n$, then $\delta \geq .13p(n-1)$ and $\Delta \leq 1.87p(n-1)$, with high probability (application of Chernoff bound). This means that $\delta^2/\Delta^4 \geq .00131/(p(n-1))$. Therefore $g \geq .39$ with high probability.

Note that the condition $p \geq 8(\ln n)/n$ is not really restrictive, because a similar condition $p \geq (1+\epsilon)(\ln n)/n$ is required for the graph to be connected with high probability.

D.2 Mixing time in random regular graphs

Friedman [10] shows that the second eigenvalue of the *adjancency matrix* of a random d-regular graph is at most $2\sqrt{d-1}+\epsilon$, with high probability, for any fix value of d. This implies our claim that the graph has constant eigenvalue gap, as explain below.

If we call A the adjacency matrix of a random d-regular graph and we call P the probability matrix of the random walk, it is easy to see that $P = \frac{1}{2}(I + A/d)$. Let λ_i (resp. μ_i) be the eigenvalues of A (resp. P). It follows that $\mu_i = \frac{1}{2}(1+\lambda_i/d)$.

The second eigenvalue is therefore $\mu_2 \leq 1/2 + \sqrt{d-1}/d + \epsilon/2d$ which is at most $0.98 + \epsilon/6$, in the worst case $d = 3$. This implies that $g = 1 - \mu_2$ is a positive constant, as needed.

E Simulation methodology

The message-level protocol simulator (written in Perl, about 1600 lines of code) chooses a node, uniformly at random, and simulates replica placement performed by that node. A search is performed from another similarly chosen node, which sends out search probes one at a time and records how many probes are necessary to find the item. This placement-search simulation is repeated for 10,000 trials with the same graph and key.

We present results from three kinds of graphs: *random*, *power-law*, and *Gnutella*. The random graphs have uniform edge probability between any two nodes. (In order to generate very large random sparse graphs, we generate an even larger sparse graph and take the giant component). We consider a variety of random graphs with sizes ranging from 10,000 to 100,000 nodes with different average degrees. In a power law graph of n nodes, node i ($i = 1, \ldots, n$) has degree $d_i = \omega/i^\alpha$ for some positive constants ω and α. The properties of these types of graphs have recently been extensively studied in the context of the Internet [20] and of peer-to-peer networks [28]. Our power-law graphs are generated using the `inet` (version 3.0) topology generator [32], which is intended to model AS-level topology of the Internet. Specifically, we consider 10,000 node `inet` graphs. For this topology size, `inet` yields graphs with average degree 4.11. The third type of graph we use represents a snapshot generated by crawling the Gnutella peer-to-peer network. The Gnutella topology has 61,274 nodes with average degree 4.70, and also exhibits some power-law type properties [28]. Nominally, for all topologies, we assume that each peer keeps identity information for 2-hop neighbors.

F Analysis of Bloom filter variant

We analyze the attenuated Bloom-filter technique for unstructured search. We assume that the Bloom filters are propagated within a h_B-hop neighborhood of each replica.

Suppose that there are r replicas placed for item i in the graph. The placement could either be uniformly at random or using local minima — the only restriction is that the h_B hop neighborhood of the replicas do not overlap too much. Let S be the union of the h_B-hop neighborhoods of the nodes where the replicas are placed, i.e. each node in S either holds a replica, or holds a Bloom filter that matches item i. On average, the number of nodes in S is around $r \cdot d^{h_B}$ (again assuming that the neighborhoods are almost distinct). There are also two ways of performing the search, when Bloom filters are used: we can do the LMS search as described in Section 3, that is to say a random walk followed by a deterministic walk, or we can just send a random walk.

Note that a search finds the item i if and only if it visits a node in the set S. Let's now consider the probability of hitting a node in the set S during the random walk: for a graph with sufficient expansion, the probability that a random walk of length l visits a node in the set S is given by results due to [13] and [17]. Specifically, Pr[search probe fails] = Pr[walk does not intersect S] is $\exp(-\Omega(l\frac{|S|}{n}))$. Note that this is a powerful result: it essentially states that there is an independent probability of $\exp(-\Omega(\frac{|S|}{n})) = \exp(-\Omega(\frac{rd^{h_B}}{n}))$ that a node in S is not visited *for every step of the random walk*.

It is possible to show that, if the LMS lookaround distance h (Section 3) is equal to the Bloom filter depth h_B, then performing the search with just a random walk of length $l+1$ is as efficient as performing a random walk of length l followed by a deterministic walk (which in general takes more than one step). The reason for this is that, after l steps of the random walk, if no element in S has been hit yet, performing an extra random step will lead to success with probability at least $1-\exp(-\Omega(\frac{rd^{h_B}}{n}))$. If *instead*, we performed a deterministic walk towards the local minimum, the probability that we succeed would be equal to the probability that there is a replica at *that* local minimum, i.e. $r/k = r \cdot d^h/n$, which is essentially the same.

Finally, we note that if the neighborhood is propagated to a larger radius than the Bloom filters, then the LMS search augmented with Bloom filter checks is more efficient than a pure random walk.

Bloom filter size estimation Assume that we are employing attenuated Bloom filters [26] of depth h_B, that each node has I items degree d. It is easy to show that, for any $p > 0$, in order to achieve a probability of false positive of at most p/d, the length of the Bloom filter m (number of bits) should satisfy:

$$m = \left(-\log\frac{p}{d}\right)(\log e) I \cdot d^{h_B-1} \tag{13}$$

This ensures that the probability that a node finds a false positive in *some* neighbor's Bloom filter is at most p.

For example, in a graph of average degree $d = 4$ with $I = 100$ items per node, if we want to achieve a false positive probability of at most $p = 10^{-5}$, we need Bloom filters with $m = 10739$ bits.