Maximizing Remote Work in Flooding-based Peer-to-Peer Systems *

Qixiang Sun Neil Daswani Hector Garcia-Molina
Computer Science Department
Stanford University, Stanford, CA 94305, USA
{qsun, daswani, hector}@cs.stanford.edu

Abstract

In peer-to-peer (P2P) systems where individual peers must cooperate to process each other's requests, a useful metric for evaluating the system is how many remote requests are serviced by each peer. In this paper we apply this remote work metric to flooding-based P2P search networks such as Gnutella. We study how to maximize the remote work in the entire network by controlling the rate of query injection at each node. In particular, we provide a simple procedure for finding the optimal rate of query injection and prove its optimality. We also show that a simple prefer-high-TTL protocol in which each peer processes only queries with the highest time-to-live (TTL) is optimal.

1 Introduction

Flooding-based peer-to-peer systems like Gnutella [4] have been deployed and used by millions of users worldwide to share and exchange files. As of April 2003, Gnutella has over one million users (with at least one hundred thousand concurrent users [5]) and ten tera-byte of shared data. Also according to [3], there are over 10 vendors actively developing Gnutella-style clients for their applications.

While there is significant research interest in distributed hash tables [8] [9] [11] [13], Gnutella-style systems are used in practice for four reasons: 1) simple to implement, 2) easy to deploy, 3) extremely robust in handling frequent peer arrivals and departures, and 4) supports wild-card searches. Moreover, in ad-hoc wireless environments where unicast is just as expensive as broadcast, a flooding-based mechanism is more desirable.

Although a flooding-based search mechanism can be inefficient as a search query is forwarded to all nodes within a certain number of hops (e.g., 7 hops), Gnutella-style networks have, nevertheless, scaled to millions of users by using a super-node architecture where high speed (CPU and bandwidth) nodes act as proxies for regular (slower) nodes. Figure 1 shows a sample super-node network with 3 super-nodes and 16 regular nodes. Each super-node indexes the content of its attached regular nodes and performs the flooding-based search on behalf of the regular nodes. In this architecture, a network with millions of users can be reduced to one with tens of thousands of super-nodes, where a flooding mechanism is adequate.

Even with this architecture, super-node networks are still susceptible to overloading when too many search queries are generated by users. In the extreme case, if every super-node uses all of its processing capacity to inject new search queries instead of answering and propagating existing queries, no "useful" work is done because queries are not answered by anyone. We define "useful" or *remote work* as a super-node processing a query that is not inject by itself or by its attached regular node. At the other extreme,

*This is the extended version of the work of the same title that appeared in DISC 2003. This version is approximately 60% longer than the original. It includes an example of oscillation, all the proofs, and a simulation result which confirms our optimal ρ selection and illustrates the relation between ρ and total remote work.

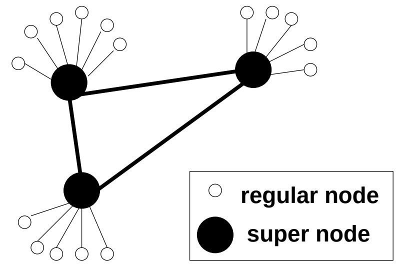


Figure 1: A sample super-node network.

if super-nodes inject too few new queries, they will have available capacity to process remote queries, but there will not be enough queries to keep the super-nodes busy. Thus, our goal is to pick a query-injection rate between these two extremes that maximizes the remote work performed.

We chose *remote work* as our objective metric because it succinctly captures the goal of users. The more remote super-nodes that process a given user query, the more potential answers the user will receive. From among the answers, the user can then select those he wants, and the larger the selection, the better. For instance, if the user searches for compositions by "Bach," he can then select titles that sound appealing, or files that have a good recording quality.

One approach to maximizing the remote work is to change the search protocol itself, e.g., using random walkers [7] or iterative deepening [12]. In this paper we attack the problem from a different angle: we control the rate of query injection at individual super-nodes. We address the following questions:

- How do we model query injection, processing, and propagation in a Gnutella-style system?
- What is the optimal number of new queries that each super-node should inject each round as to maximize the remote work done in a network?
- What is the impact of using different protocols to select which queries to process and propagate? Is there an optimal protocol?
- Should we enforce a fair policy where every super-node injects the same number of new queries into the network? Or should highly-connected super-nodes in the "critical" part of the network inject more queries (or less)?
- What is the penalty in terms of reduced remote work for using a fair policy?
- What are some heuristics for more complex systems that are outside of our simple model?

Daswani et al. in [1] conducted simulations to answer some of the above questions focusing on the impact of malicious super-nodes who purposely generate large number of bogus queries to reduce the amount of "useful" work done in a flooding-based peer-to-peer system. In the current paper, we do not consider malicious super-nodes doing denial-of-service (DoS) attacks using bogus queries. Instead, we assume all super-nodes are cooperating to maximize useful work in the network. The results in this paper provide a firm theoretical foundation for studying the effects of DoS attacks and establish a baseline of comparison. These results can be easily incorporated into [1] to further extend their results.

Knowing the theoretical optimal rate of query injection and the maximum remote work possible can improve the construction of the overlay network. For example, a super-node can use the optimal query-injection rate to dynamically decide whether it should accept more clients or disconnect existing ones. We can also use remote work as metric to evaluate different types of overlay topologies.

Although our work is specific to Gnutella-like systems, we do address an issue that we believe will be of growing importance in distributed systems, that of getting autonomous components to provide services for each other. Whether the system is a publish-subscribe one, or a sensor net, or an ad-hoc wireless network, nodes must balance their local needs (e.g., disseminate events or messages originating locally) with the services they provide to others (e.g., packet forwarding, resource discovery). As far as we know, this "distributed resource coordination" problem has not been studied in detail. Our paper is a first study of such coordination for autonomous systems.

2 Assumptions and a Model

We use a very simple model of Gnutella to capture key performance characteristics that are relevant to our goal of maximizing remote work. Given that regular nodes always access the network via a super-node, we only need to capture the activities of the super-nodes. Specifically, we model the super-node network as a graph $G = (V, E)$ where edges represent connections between super-nodes. For brevity, when we say "node" in the remainder of this paper, we mean super-node unless stated otherwise explicitly.

We model the P2P system as operating in rounds, where search queries are injected and processed during the round and forwarded to neighboring peers between rounds. Although the system does not have to be sychronous, we will assume sychrony for analysis purposes. Note that queries "injected" by a super-node are typically initiated by the regular nodes attached to it.

We assume each query has a time-to-live (TTL) field that is decremented by one each time when forwarded to other peers. When the TTL becomes negative, the query is removed from the network. For our purpose of maximizing remote work, we only model the propagation of search queries and ignore other communication such as search replies, ping-pong messages, and actual file transfers.

We also assume the bottleneck of the system is the processing capacity of the super-nodes rather than the network bandwidth. Furthermore, we assume receiving queries from the network has negligible processing cost as compared to the actual processing of a query. There are three reasons for these assumptions: (1) super-nodes have excellent network connectivities, e.g., 10 megabits or better; (2) backbone bandwidth is grossly over-provisioned; and (3) wild-card search queries are expensive to evaluate because simple hashing techniques do not work well.

We assign each super-node a processing capacity of C queries per round. A super-node may use its capacity in two ways: (1) accept and process a new search query from an attached regular node, or (2) process a remote query forwarded to it by a neighboring super-node. We refer to case 1 as a super-node *injecting new queries*, and refer to case 2 as *processing remote queries*. For clarification, processing a remote query involves two steps: one, match the query against the shared data indexed by this super-node; and two, forward this query to neighboring nodes. Obviously in a single round, the number of new queries injected plus the number of remote queries processed is at most C.

In most of our analysis in this paper, we assume all nodes have the same processing capacity to make the analysis tractable. Although Sariou et. al. [10] observed large variations among Gnutella clients, variations among super-nodes are much smaller. We will briefly outline the difficulties in handling super-nodes with different capacities as an open problem in Section 9.2.

Although each node can only process up to C queries per round, its neighboring nodes may send it more than C remote queries. Because we assumed that network bandwidth is not the limiting factor and that the cost of receiving data from the network is negligible, we allow each node to receive all the incoming remote queries even if it does not have the capacity to process them all. A node must then decide which remote queries to process this round and drop the remaining queries. We do not allow a node to "temporarily" buffer excess *remote* queries for processing at a later round because we are interested in the long-term system behavior where nodes are constantly overloaded.

The long-term behavior of a peer-to-peer system certainly depends heavily on how each node decides which queries to process and drop. For brevity, we use the term *protocol* to refer to a node's decision mechanism. As an example, a node is said to be using a random protocol if it picks which queries to process uniformly at random.

One important parameter of a protocol is how a node divides its capacity between injecting new queries and processing remote queries. We use a fraction ρ between 0 and 1 to denote this parameter. For example, $\rho = \frac{1}{3}$ implies one third of a node's capacity is allocated for injecting new queries while the other two third is used for processing remote queries. We assume that a super-node injects its full quota of ρC new queries each round, i.e., there is always an abundance of queries that regular nodes want to submit. This assumption is reasonable because our goal is to study the maximum amount of remote work possible which can only occur if nodes are generating sufficient number of new queries to keep the system busy. In practice, a super-node can inject new *local* queries at a fixed rate by buffering and delaying new search queries from its attached regular nodes.

Rather than trying to build an accurate model that can predict the actual performance of the peer-to-peer system, we have made many simplifying assumptions to make our study of the fundamental system behavior feasible. This simplified model retains all the important aspects of a flooding-based peer-to-peer protocol and does not restrict design decisions.

3 Notation and Problem Definition

- ρ_v denotes the fraction of processing capacity node v allocates for injecting new queries per round.
- $\bar{\rho} = \{\rho_v \mid v \in V\}$ denotes the set of ρ_v used by all nodes in network G.
- $\delta(u, v)$ denotes the minimum hop distance between nodes u and v in network G.
- $D(v, \tau)$ denotes the set of nodes u, excluding v, in G such that $\delta(u, v) \leq \tau$.
- $\hat{D}(v, \tau)$ denotes $D(v, \tau) \cup \{v\}$.
- $W_t^{\mathcal{P}}(v, \bar{\rho})$ denotes the set of queries processed by node v, using protocol $\mathcal{P}$ with settings $\bar{\rho}$ for the nodes, during round t. The set $W_t^{\mathcal{P}}(v, \bar{\rho})$ includes both new queries injected by v and processed remote queries. We drop the superscript $\mathcal{P}$ when the context is clear.
- $R_t^{\mathcal{P}}(v, \bar{\rho}) \subset W_t^{\mathcal{P}}(v, \bar{\rho})$ denotes the set of remote queries processed by node v at time t.
- $RW_t^{\mathcal{P}}(\bar{\rho}) = \sum_{v \in V} |R_t^{\mathcal{P}}(v, \bar{\rho})|$ denotes the number of remote queries processed by all nodes in network G at time t.

With the notation above, maximizing the remote work of a network G using protocol $\mathcal{P}$ can be stated formally as:

Problem: Given a graph $G = (V, E)$, maximum TTL τ, processing capacity C, and a protocol $\mathcal{P}$, find the optimal rate of injecting new queries $\bar{\rho} = \{\rho_v \mid v \in V\}$ such that $\sum_t RW_t^{\mathcal{P}}(\bar{\rho})$ is maximized.

The maximization problem is stated above as the cumulative number of remote queries processed over all nodes and all time. We chose to sum over all time to take into account of protocols with nondeterministic or irregular behaviors. However, as we will see, the protocols studied here all have some form of "steady-state" behavior.

4 Protocols

Before describing the protocols, we first need to discuss how to tag each query with an ID to avoid processing duplicate queries and to remove queries when their TTL expires. For a query q, we use a triplet

Deterministic Prefer-High-TTL Protocol $\mathcal{H}^{\mathcal{D}}$

During every round, each node $v \in V$ performs the following tasks in the order shown below:

1. Inject $\rho_v \cdot C$ new queries with the triplet identifiers $\{v, \tau, 1\}, \{v, \tau, 2\}, \ldots, \{v, \tau, \rho_v C\}$. Denote this set of local queries L_v. (For clarity in the presentation, we assume $\rho_v C$ is an integer. We can take the floor if it is not an integer.)
2. Sort all incoming queries from adjacent super-nodes in decreasing order of TTL, break ties in a deterministic manner that is independent of the current time, and remove queries that are duplicates or have already been processed at some previous time step. Denote this sorted list of new incoming queries I_v.
3. Take the first $(1 - \rho_v)C$ queries in I_v. Denote this set of remote queries R_v.
4. Service queries in L_v and R_v against local index.
5. Decrement the TTL of queries in L_v and R_v by 1.
6. Forward all queries in L_v and R_v that have $TTL \geq 0$ to all neighbors.

Figure 2: An informal description of the deterministic prefer-high-TTL protocol.

(src, ttl, mid) where src is the node that injected the query, ttl is the current time-to-live of q as q moves around the network, and mid is an internal sequence number where $1 \leq mid \leq C$. We enforce three invariants about the IDs: (1) for any two queries injected by the same node in the same round, their mids are different; (2) $0 \leq ttl \leq \tau$ where τ is the maximum TTL; and (3) a query with ID (src, ttl, mid) at time t is injected at time $t - \tau + ttl$.

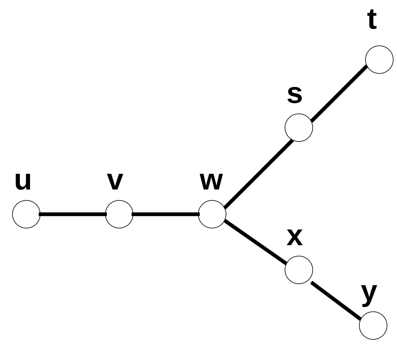
Note that when the query travels around the network, its ID changes as the ttl is decremented. To determine whether two query IDs q_1 and q_2 at times t_1 and t_2, respectively, refer to the same query, we check whether these two IDs have the same src node, the same mid, and were injected into the network at the same time. For example, assuming all queries initially have a TTL τ when injected, then a query with ID $q_1 = (u, 5, 2)$ at time step 8 is the same query as a query with ID $q_2 = (u, 3, 2)$ at time step 10 because both queries are injected by node u at time $8 + 5 - \tau = 10 + 3 - \tau = 13 - \tau$ with sequence number 2.

Using these IDs, we describe the operations of the deterministic prefer-high-TTL protocol $\mathcal{H}^{\mathcal{D}}$ in Figure 2. Essentially, after each node injects its new queries for the round, it then processes remote queries in decreasing TTL order until the processing capacity has been exhausted. If two queries have the same TTL, the tie is broken deterministically, e.g., lexicographically by source node ID and then the sequence number.

Similarly, the randomized prefer-high-TTL protocol $\mathcal{H}^{\mathcal{R}}$ performs the same steps as $\mathcal{H}^{\mathcal{D}}$ except ties are broken randomly. Though $\mathcal{H}^{\mathcal{R}}$ and $\mathcal{H}^{\mathcal{D}}$ are very similar, they exhibit different steady-state behavior as we will see in the next section. This distinction has significant impact on how efficiently we can simulate the protocols for experimental studies. A third protocol that we will use for illustrative purposes is the prefer-low-TTL protocol $\mathcal{L}$. Instead of sorting all the incoming queries in the set I_v in decreasing order of TTL during step 2 (of Figure 2), protocol $\mathcal{L}$ sorts the queries in increasing order of TTL.

5 Steady State

Regardless of the transient behavior at the beginning of time, a protocol that processes the most remote queries in the steady state will process the most remote work in the long run. Therefore, if two protocols have steady states, then we can simply compare their per-round performance in the steady state. It turns out that not all protocols have some form of steady state. To illustrate, consider an example topology consisting of seven nodes shown in Figure 3. If we use the prefer-low-TTL protocol $\mathcal{L}$ with maximum TTL $\tau = 2$ and $\rho_v = \frac{1}{4}$ for all nodes v (i.e., each node injects $\frac{C}{4}$ new queries per round), then we observe an oscillation.



$\tau = 2 \qquad \rho = \frac{1}{4}$

Figure 3: An example where prefer-low-TTL does not have a steady state.

Time	w	u	v	x	y	s	t
0	$\frac{1}{4}w_0$	$\frac{1}{4}u_0$	$\frac{1}{4}v_0$	$\frac{1}{4}x_0$	$\frac{1}{4}y_0$	$\frac{1}{4}s_0$	$\frac{1}{4}t_0$
1	$\frac{1}{4}w_1, \frac{1}{4}v_0,$ $\frac{1}{4}x_0, \frac{1}{4}s_0$	$\frac{1}{4}u_1, \frac{1}{4}v_0$	$\frac{1}{4}v_1, \frac{1}{4}u_0,$ $\frac{1}{4}w_0$	$\frac{1}{4}x_1, \frac{1}{4}w_0,$ $\frac{1}{4}y_0$	$\frac{1}{4}y_1, \frac{1}{4}x_0$	$\frac{1}{4}s_1, \frac{1}{4}w_0,$ $\frac{1}{4}t_0$	$\frac{1}{4}t_1, \frac{1}{4}s_0$
2	$\frac{1}{4}w_2, \frac{1}{4}u_0$ $\frac{1}{4}y_0, \frac{1}{4}t_0$	$\frac{1}{4}u_2, \frac{1}{4}w_0$ $\frac{1}{4}v_1$	$\frac{1}{4}v_2, \frac{1}{4}x_0$ $\frac{1}{4}s_0, \frac{1}{4}w_1$	$\frac{1}{4}x_2, \frac{1}{4}v_0$ $\frac{1}{4}s_0, \frac{1}{4}w_1$	$\frac{1}{4}y_2, \frac{1}{4}w_0$ $\frac{1}{4}x_1$	$\frac{1}{4}s_2, \frac{1}{4}v_0$ $\frac{1}{4}x_0, \frac{1}{4}w_1$	$\frac{1}{4}t_2, \frac{1}{4}w_0$ $\frac{1}{4}s_1$
3	$\frac{1}{4}w_3, \frac{1}{4}v_2$ $\frac{1}{4}x_2, \frac{1}{4}s_2$	$\frac{1}{4}u_3, \frac{1}{4}w_1$ $\frac{1}{4}v_2$	$\frac{1}{4}v_3, \frac{1}{4}w_2$ $\frac{1}{4}u_2$	$\frac{1}{4}x_3, \frac{1}{4}w_2$ $\frac{1}{4}y_2$	$\frac{1}{4}y_3, \frac{1}{4}w_1$ $\frac{1}{4}x_2$	$\frac{1}{4}s_3, \frac{1}{4}w_2$ $\frac{1}{4}t_2$	$\frac{1}{4}t_3, \frac{1}{4}w_1$ $\frac{1}{4}s_2$
4	$\frac{1}{4}w_4, \frac{1}{4}v_2$ $\frac{1}{4}x_2, \frac{1}{4}t_2$	$\frac{1}{4}u_4, \frac{1}{4}w_2$ $\frac{1}{4}v_3$	$\frac{1}{4}v_4, \frac{1}{4}x_2$ $\frac{1}{4}s_2, \frac{1}{4}w_3$	$\frac{1}{4}x_4, \frac{1}{4}v_2$ $\frac{1}{4}s_2, \frac{1}{4}w_3$	$\frac{1}{4}y_4, \frac{1}{4}w_2$ $\frac{1}{4}x_3$	$\frac{1}{4}s_4, \frac{1}{4}v_2$ $\frac{1}{4}x_2, \frac{1}{4}w_3$	$\frac{1}{4}t_4, \frac{1}{4}w_2$ $\frac{1}{4}s_3$

Table 1: Execution trace of protocol $\mathcal{L}$ on the example in Figure 3. Note v and x oscillate.

Table 1 traces out which queries each node processes during the first four rounds when executing protocol $\mathcal{L}$.

In Table 1 we use the notation αz_i to indicate a node has processed queries that were injected by node z at time i with internal sequence numbers 1 through αC. For instance, at time 0, all nodes are only processing queries injected by themselves. At time 1, each node is processing its own newly injected queries and remote queries injected by its neighbors at time 0.

In general when tracing out which queries are processed by node z at time i, we simply look at the set of queries Q that are processed by z's neighbors at time $i-1$. We first eliminate queries in Q whose TTL have expired or have already been processed before. We then sort in increasing TTL order the remaining queries in Q where ties are broken by the source node ID. (Specifically, we break ties in the following order: w, v, s, x, u, t, and then y.) For example, consider the entry in Table 1 that corresponds to node w at time 2. To fill in the square, we form the set Q from queries processed by nodes v, x, and s at time 1, which is $\{\frac{1}{4}v_1, \frac{1}{4}u_0, \frac{1}{4}w_0, \frac{1}{4}x_1, \frac{1}{4}y_0, \frac{1}{4}s_1, \frac{1}{4}t_0\}$. We first filter out $\frac{1}{4}w_0$ because we have handled it in the previous time step. We then sort the remaining queries according to TTL to get $\{\frac{1}{4}u_0, \frac{1}{4}y_0, \frac{1}{4}t_0, \frac{1}{4}v_1, \frac{1}{4}x_1, \frac{1}{4}s_1\}$. Because node w only has $\frac{3}{4}$ capacity left after its own queries, w simply selects $\{\frac{1}{4}u_0, \frac{1}{4}y_0, \frac{1}{4}t_0\}$ to fill its capacity.

The key point to notice in this trace is that node w decided to drop queries from its immediate neighbors at time step 2 due to the low TTL preference. As a result, when we progress to time step 3, node w is only able to forward the queries generated by itself because the TTL on the remaining queries have expired. This lack of forwarded queries causes w's neighbors to process fewer queries than the previous time step.

If we continue the trace to time step 4, it is identical to time step 2, with time indices advanced by 2. Hence tracing the executing further will yield an oscillation between steps 2 and 3. The cause of the oscillation is due to node w acting as a choke point for forwarding queries on the even-numbered time steps. For this simple example, we could have foreseen this period 2 osicllation. Unfortunately, in general, protocol $\mathcal{L}$'s period of oscillation is a function of τ, the tie breaking policy, and the network topology G, thus not easy to determine a priori.

Oscillations aside, there are two flavors of steady state that are of particular interest because they distinguish between protocols $\mathcal{H}^{\mathcal{D}}$ and $\mathcal{H}^{\mathcal{R}}$. The first kind is a *strong* steady state where we can determine exactly which queries will be processed by every node. Formally,

Definition 1. *(Strong steady state) A protocol $\mathcal{P}$ has a* strong steady state *if given any $\bar{\rho}$, there exists t_0 such that for every node v and all $t > t_0$, $R_t^{\mathcal{P}}(v, \bar{\rho}) = R_{t_0}^{\mathcal{P}}(v, \bar{\rho})$.*

In other words, *strong steady state* guarantees that after time t_0, each node will process remote queries with the same triplet ID as the previous time step. For example, if node v processed a query with ID $(u, 5, 2)$ at time t_0, then v will process a query of the same ID from then on. Thus having a strong steady state makes simulation studies easier. Note that the same triple ID at two different times does not mean the same query because the two queries are created at different times.

An alternative is to relax the constraint of processing queries with the same triplet IDs.

Definition 2. *(Weak steady state) A protocol $\mathcal{P}$ has a* weak steady state *if given any $\bar{\rho}$, there exists t_0 such that for every node v and all $t > t_0$, $|R_t^{\mathcal{P}}(v, \bar{\rho})| = |R_{t_0}^{\mathcal{P}}(v, \bar{\rho})|$.*

A weak steady state only requires the number of remote queries processed to be the same rather than the query IDs to be the same. Since our objective is to maximize the total number of remote queries processed, having a weak steady state is sufficient for our analysis. Clearly, strong steady state implies weak steady state.

With these two notions of steady state, we now show protocol $\mathcal{H}^{\mathcal{D}}$ has a strong steady state. In particular, we show $\mathcal{H}^{\mathcal{D}}$ has a monotonicity property.

Proposition 3. *(Monotonicity) In protocol $\mathcal{H}^{\mathcal{D}}$, given $\bar{\rho}$, for any node v and a query ID $q = (src, ttl, mid)$,*

1. *if $q \in W_{\tau - ttl_q}(v, \bar{\rho})$, then $q \in W_t(v, \bar{\rho})$ for all $t \geq \tau - ttl_q$.*
2. *if $q \in W_t(v, \bar{\rho})$ for some $t > \tau - ttl_q$, then $q \in W_{\tau - ttl_q}(v, \bar{\rho})$.*

Informally, monotonicity states that once a query ID q is in $W_{t_1}(v, \bar{\rho})$ for any node v and time t_1, the ID q can never disappear from $W_{t_2}(v, \bar{\rho})$ for all $t_2 > t_1$. It also guarantees the first appearance of q is at time $\tau - ttl_q$. The monotonicity is the result of breaking ties among queries of the same TTL in a deterministic fashion. Before proving this claim, we first note that in handling the set of queries W_t, because the TTL of a query is decremented during a round, there are two possible TTLs for each query $q \in W_t(v, \bar{\rho})$. For consistency, we use the TTL before the decrement as the TTL of query q. We now give the formal proof.

Proof. We prove our claim by induction on the number of hops a query has traveled, which is $(\tau - ttl)$.

Base Case: a query q with $\tau - ttl_q = 0$. This case occurs when $ttl_q = \tau$, i.e., q is a local query that was just created. Since protocol $\mathcal{H}^{\mathcal{D}}$ injects the same number of local queries with the same message IDs 1 through $\rho_v \cdot C$ for each node v at each time step, our claim holds trivially.

Inductive Step: Assume our claim holds for all queries q that have traveled less than η hops, i.e., with $\tau - ttl_q < \eta$, we want to show that our claim also holds for queries with $\tau - ttl_q = \eta$. For part (1), assume $q = (src, \tau - \eta, mid) \in W_\eta(v, \bar{\rho})$ for some node v. Now we need to show $q \in W_t(v, \bar{\rho})$ for all $t \geq \eta$,.

Consider the set of queries Q with TTL = $\tau - \eta$ that is processed by node v at time t and the set of queries P with TTL = $\tau - \eta + 1$ that is processed by all neighbors of node v at time $t-1$. Notice that $Q \subset P'$ where $P' = \{(src, ttl, mid) \mid (src, ttl+1, mid) \in P\}$. By our induction hypothesis and using part (2) of the claim, the set P is the same for $t \geq \eta - 1$. Thus node v receives the same set P' for all $t \geq \eta$. By construction, protocol $\mathcal{H}^{\mathcal{D}}$ deterministically selects the same subset Q from P' for all $t \geq \eta$. Therefore, if $q \in Q$ at time η, then $q \in Q$ for all $t \geq \eta$, as required. Part (2) of the claim is similar. □

The monotonicity property can be used directly to show that $\mathcal{H}^{\mathcal{D}}$has a steady state.

Theorem 4. *Protocol $\mathcal{H}^{\mathcal{D}}$ reaches a strong steady state in τ time steps.*

Proof. For any node v, let $L_t(v, \bar{\rho}) = \{q \mid q \in W_t(v, \bar{\rho}) \text{ and } src_q = v\}$ denote the set of locally injected queries. Then $R_t(v, \bar{\rho}) = W_t(v, \bar{\rho}) - L_t(v, \bar{\rho})$. Since $L_t(v, \bar{\rho})$ is constant for all t, to prove our claim of achieving strong steady state in τ steps, it is sufficient to show that for every node v and all $t > \tau$, $W_t(v, \bar{\rho}) = W_\tau(v, \bar{\rho})$; specifically, $W_t(v, \bar{\rho}) \subset W_\tau(v, \bar{\rho})$ and $W_\tau(v, \bar{\rho}) \subset W_t(v, \bar{\rho})$.

For any $q \in W_t(v, \bar{\rho})$, using part (2) of Proposition 3, $q \in W_{\tau - ttl_q}(v, \bar{\rho})$. Because $ttl_q \geq 0, \tau - ttl_q \leq \tau$. Applying part (1) of Proposition 3, we get $q \in W_\tau(v, \bar{\rho})$, which implies $W_t(v, \bar{\rho}) \subset W_\tau(v, \bar{\rho})$. Similarly, $W_\tau(v, \bar{\rho}) \subset W_t(v, \bar{\rho})$. □

Unlike protocol $\mathcal{H}^{\mathcal{D}}$, the randomized version $\mathcal{H}^{\mathcal{R}}$ only has a weak steady state. Clearly $\mathcal{H}^{\mathcal{R}}$ does not have a strong steady state because the random selections do not guarantee a node will consistently choose remote queries with the same IDs. The fact that $\mathcal{H}^{\mathcal{R}}$ has a weak steady state is a directly corollary of a theorem in the next section that states both protocols $\mathcal{H}^{\mathcal{D}}$ and $\mathcal{H}^{\mathcal{R}}$ are "optimal" in the number of remote queries processed. Since $\mathcal{H}^{\mathcal{D}}$ and $\mathcal{H}^{\mathcal{R}}$ processes the same number of remote queries and $\mathcal{H}^{\mathcal{D}}$ reaches a strong steady state in τ time steps, then $\mathcal{H}^{\mathcal{R}}$ must reach a weak steady state in τ time steps.

6 Optimality of Protocol $\mathcal{H}^{\mathcal{R}}$

We now show that for any settings of $\bar{\rho}$, the two prefer-high-TTL protocols, $\mathcal{H}^{\mathcal{D}}$ and $\mathcal{H}^{\mathcal{R}}$, processes as much remote work as any other protocols using the same $\bar{\rho}$ settings, and hence are optimal. Since protocol $\mathcal{H}^{\mathcal{D}}$ is a special case of protocol $\mathcal{H}^{\mathcal{R}}$, we only show the optimality of protocol $\mathcal{H}^{\mathcal{R}}$. We prove this claim by first establishing an upper bound on the amount of remote work any protocol can process, and then showing protocol $\mathcal{H}^{\mathcal{R}}$achieves this upper bound.

For the upper bound, notice that regardless of which protocol we use, the number of remote queries a node v can process, $|R_t(v, \bar{\rho})|$, is limited by two factors: (1) node v's processing capacity, and (2) how many queries are injected by nodes within τ hops of v. At maximum capacity, a node v can process $(1 - \rho_v)C$ queries per round. We call such a node *saturated*.

When a node v is not saturated, it can receive up to $K_v = C \cdot \sum_{u \in D(v,\tau)} \rho_u$ queries from nodes within τ hops. For protocols without steady state, the actual number of queries processed by node v may vary between rounds, (e.g., process no queries during one round, but a large amount the next round); however, the average number of queries processed per round, over time, is bounded by K_v.

We get our upper bound by combining the two limiting factors and taking the minimum number of remote queries processed in case 1 and case 2 (along with a special case when $t < \tau$).

Proposition 5. *For any protocol $\mathcal{P}$, any node v, and any setting $\bar{\rho}$,*

$$\sum_t |R_t(v, \bar{\rho})| \leq C \cdot \sum_t \min \left(1 - \rho_v, \sum_{w \in D(v, \min(\tau, t))} \rho_w \right)$$

We now show in two steps that protocol $\mathcal{H}^{\mathcal{R}}$achieves this upper bound. In the first step, we claim that if a node v's "neighbors" cannot inject enough queries to continuously saturate v, then node v will process every query injected by these "neighbors." Stated formally,

Lemma 6. *Consider protocol $\mathcal{H}^{\mathcal{R}}$ and any node v. Suppose for some hop count $h \le \tau$, $\sum_{w \in D(v,h)} \rho_w \le 1 - \rho_v$. Then for all nodes $w \in D(v,h)$ and all i such that $1 \le i \le \rho_w C$, the query with triplet ID $(w, \tau - \delta(w,v), i) \in R_t(v, \bar{\rho})$ for all time $t \ge \delta(w,v)$.*

Proof. Note by assuming $\sum_{w \in D(v,h)} \rho_w \le 1 - \rho_v$, no queries are dropped due to lack of capacity, i.e., there are no random choices in deciding which queries to drop. Therefore, this lemma becomes a special case of the monotonicity property in Proposition 3. The same induction proof holds here. □

In the second step, we claim that if node v's "neighbors" are continuously injecting more queries than v can process, then node v processes exactly $(1 - \rho_v)C$ queries each round. Formally,

Lemma 7. *In protocol $\mathcal{H}^{\mathcal{R}}$, for any node v and hop count h, if $\sum_{w \in D(v,h)} \rho_w > 1 - \rho_v$, then node v is saturated after time h, i.e., $|R_t(v, \bar{\rho})| = (1 - \rho_v)C$ for all $t \ge h$.*

This claim is not immediately obvious because the random selections in protocol $\mathcal{H}^{\mathcal{R}}$ may result in many duplicate queries arriving at a node v and reduce the number of remote queries processed. Fortunately, the prefer-high-TTL mechanism ensures "enough" non-duplicate queries arrive at v to saturate its processing capacity. We now give the formal proof.

Proof. Define σ_v to be the smallest hop count where $\sum_{w \in D(v,\sigma_v)} \rho_w > 1 - \rho_v$. In other words, nodes that are less than σ_v hops away from v cannot inject enough queries to saturate v. We show our claim by induction on σ_v.

Base Case: $\sigma_v = 1$. This case corresponds to the situation where node v's immediate neighbors, denoted by $N(v)$, are injecting more queries than v can handle, i.e., $\sum_{w \in N(v)} \rho_w > 1 - \rho_v$. Since nodes in $N(v)$ forward newly injected queries to v each round, node v receives at least $(1 - \rho_v)C$ new queries each round. Thus node v must be processing at maximum capacity, or $|R_t(v, \bar{\rho})| = (1 - \rho_v)C$.

Induction Step: Assuming the claim holds for all nodes v where $\sigma_v = h$, we show the claim also holds for all nodes v where $\sigma_v = h+1$. Suppose $\sigma_v = h+1$ for some node v. Consider the immediate neighboring nodes $N(v)$. There are two cases for $N(v)$: (1) for all $w \in N(v)$, $\sum_{u \in D(w,h)} \rho_u \le 1 - \rho_w$, i.e., $\sigma_w > h$; and (2) there exists $w \in N(v)$ such that $\sum_{u \in D(w,h)} \rho_u > 1 - \rho_w$.

In case 1, note that every node that is $h+1$ hops away from v is h hops away from some node in $N(v)$. Because $\sigma_w > h$ for all $w \in N(v)$, we can apply Lemma 6 to see that all queries injected by nodes that are exactly $h+1$ hops away from v will be processed by some node in $N(v)$ after time h and forwarded to node v. Consequently, for all time $t \ge h+1$, all queries from $h+1$ hops away will reach v via nodes in $N(v)$. These queries are not duplicates of old queries because they traveled along the shortest path. Hence if $\sum_{w \in D(v,\sigma_v = h+1)} \rho_w > 1 - \rho_v$, then node v receives at least $(1 - \rho_v)C$ new queries each round, i.e., $|R_t(v, \bar{\rho})| = (1 - \rho_v)C$ for all $t \ge h+1$.

In case 2, let $w \in N(v)$ be the node where $\sum_{u \in D(w,h)} \rho_u > 1 - \rho_w$. Intuitively, some of the queries processed by node w are injected by nodes that are $h+1$ hops away from v. Call this set of queries Q. We show that $|Q|$ is sufficiently large when combined with queries injected by nodes within h hops of v, denoted by P, to saturate v. Note that $\sigma_w = h$ because if $\sigma_w < h$, then σ_v is at most h.

Applying our induction hypothesis for node w, we know w is saturated for all time $t \ge h$. Now consider the set of nodes $X = \{u | u \ne v, \delta(u,v) \le h\}$ and $Y = \{u | u \ne w, \delta(u,w) \le h\}$. Let $Z = X \cap Y$ and $U = Y - X$. Notice that nodes in U are exactly $h+1$ hops away from v and that Q is the set of queries injected by nodes in U.

Because at most $C \cdot \sum_{u \in Z} \rho_u$ queries did not originate from some node in U, we get $|Q| \geq C \cdot (1 - \rho_w - \sum_{u \in Z} \rho_u)$. Using the fact that $(Z \cup \{w\}) \subset (X \cup \{v\})$, we know $\sum_{u \in Z} \rho_u \leq \sum_{u \in X} \rho_u + \rho_v - \rho_w$. Therefore, $|Q| \geq C \cdot (1 - \rho_w - \sum_{u \in Z} \rho_u) \geq C \cdot (1 - \rho_v - \sum_{u \in X} \rho_u)$.

Also because $\sigma_v = h + 1$, Lemma 6 says node v will receive all queries from nodes within h hops of v. Hence we get that v processed $|P| = C \cdot \sum_{u \in X} \rho_u$ queries from nodes within h hops. Combining P and Q, the amount of remote work done per round $|R_t(v, \bar{\rho})|$ is at least $|P| + |Q| \geq C \cdot (\sum_{u \in X} \rho_u + 1 - \rho_v - \sum u \in X \rho_u) = (1 - \rho_v)C$, which proves our claim. □

Combining Lemmas 6 and 7 with $h = \tau$, we get that if node v's neighbors within τ hops do not inject enough queries to saturate v's processing capacity, then node v processes every query injected by them. On the other hand, if there is more than enough queries, then node v processes at maximum capacity $(1 - \rho_v)C$. Consequently,

Theorem 8. *In protocol $\mathcal{H}^{\mathcal{R}}$, for any node v and any setting $\bar{\rho}$,*

$$|R_t(v, \bar{\rho})| = C \cdot \min\left(1 - \rho_v, \sum_{u \in D(v, \min(\tau, t))} \rho_u\right)$$

By applying Theorem 8, we obtain that $\sum_t |R_t(v, \bar{\rho})|$ is equal to the upper bound established in Proposition 5. There are two immediate consequences of this observation:

Corollary 9. *Protocol $\mathcal{H}^{\mathcal{R}}$ has a weak steady-state after τ time steps.*

Corollary 10. *No protocols can achieve more remote work than protocol $\mathcal{H}^{\mathcal{R}}$.*

Corollary 10 gives us our claim that protocol $\mathcal{H}^{\mathcal{R}}$ is optimal.

Another important consequence of Theorem 8 is that in computing remote work, we do not have to worry about which queries were duplicates or which path a query traveled on. Therefore, we can treat all queries as indistinguishable from each other and rewrite our optimization problem into a simple linear program (LP). Specifically, let r_v denote the number of remote queries processed by node v in the steady state of $\mathcal{H}^{\mathcal{R}}$. Then maximizing remote work is equivalent to the objective function

$$max : \sum_v r_v \tag{1}$$

The constraints of this linear program are the two terms in the minimum clause of Theorem 8, i.e., nodes may not exceed their processing capacity (Eq. 2) and may not process more remote work than is injected by their neighbors (Eq. 3). More formally,

$$\begin{aligned} r_v &\leq C(1 - \rho_v) \quad \forall\, v \in V &(2)\\ r_v &\leq C \cdot \sum_{w \in D(v, \tau)} \rho_w \quad \forall\, v \in V &(3) \end{aligned}$$

Because the optimal r_v solutions from the linear program will be tight (i.e., equality) for either constraints 2 or 3, it is identical to taking the minimum of the two constraints. Therefore, the sum of the r_v's is precisely the number of remote queries processed in the network per round.

Unfortunately, solving the LP gives us little insight into the problem's structure. The next section builds such insights for a special case of the problem where each node has the same ρ setting, i.e., $\rho_v = \rho$ for all v.

Find_Optimal_Single_ρ:

1. order the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ such that $|\hat{D}(v_i, \tau)| \leq |\hat{D}(v_{i+1}, \tau)|$.
2. construct the sequence of non-increasing real numbers $\{d_1, d_2, \ldots d_n\}$ where $d_i = \frac{1}{|\hat{D}(v_i,\tau)|}$.
3. find the smallest k such that $\sum_{i=1}^{k} |\hat{D}(v_i, \tau)| \geq n$.
4. return d_i.

Figure 4: Procedure for finding the optimal $\hat{\rho}$ when all nodes have the same ρ.

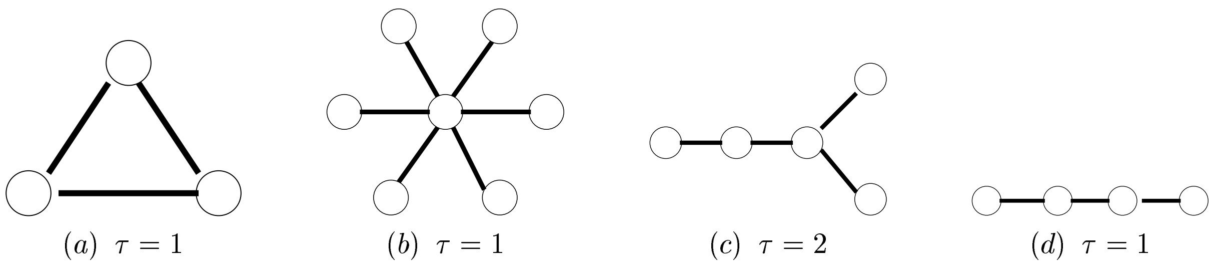


Figure 5: Four example topologies.

7 Identical ρ for All Nodes

The instance of every node having the same ρ is of particular interest because it captures fairness in the super-node network. In other words, every super-node injects the same number of new queries into the network. This instance also arises when the software clients have a hard-coded and pre-determined capacity allocation. Clearly, finding the optimal $\hat{\rho}$ setting that maximizes the total remote work is dependent on the network topology. In addition to presenting a procedure for selecting the optimal $\hat{\rho}$, we also show that imposing this "fair" criterion of identical ρ for all nodes does not significantly reduce the maximum amount of remote work.

Figure 4 shows our procedure for selecting $\hat{\rho}$. To illustrate, consider examples (a), (b), and (c) in Figure 5. We first write out $|\hat{D}(v_i, \tau)|$ for all nodes v_i in a non-decreasing sequence, and then add the numbers in sequence from the beginning until the sum exceeds the number of nodes. When we stopped adding at node i, the optimal $\hat{\rho}$ is the corresponding $d_i = \frac{1}{|\hat{D}(v_i,\tau)|}$. In example (a), we get the sequence of $|\hat{D}(v_i, \tau)|$ as $\{3, 3, 3\}$. Because 3 is the number nodes in this network, we stop immediately at $i = 1$ and get the optimal $\hat{\rho} = \frac{1}{3}$, as expected. Moving to the more complicated examples, we see example (b) generates the sequence $\{2, 2, 2, 2, 2, 2, 7\}$. After adding the first four 2s, we get $8 > 7$, thus the optimal $\hat{\rho} = d_4 = \frac{1}{|\hat{D}(v_4,\tau)|} = \frac{1}{2}$. In example (c), we get the sequence $\{3, 4, 4, 5, 5\}$ which yields the optimal $\hat{\rho} = \frac{1}{4}$ when $3 + 4 > 5$.

Before we formally prove the correctness of the *Find_Optimal_Single_ρ* procedure, we first outline the general idea behind the proof. Note that given any ρ, we can divide the nodes into two categories: the set of saturated nodes S and the set of unsaturated nodes U. Now consider using $\rho' = \rho + \epsilon$ for some $\epsilon > 0$. For all nodes $v \in S$, v's remote work is reduced by ϵ, i.e., we lose a total of $R^- = \epsilon|S|$. However, for all nodes $w \in U$, w's remote work has increased by $\epsilon|\hat{D}(w, \tau)|$, or we gain $R^+ = \epsilon \sum_{w \in U} |\hat{D}(w, \tau)|$. Thus intuitively, when $R^- = R^+$, we have found a candidate for the optimal $\hat{\rho}$. Fortunately, there is only one such candidate, which our *Find_Optimal_Single_ρ* procedure finds.

We now prove the correctness of the *Find_Optimal_Single_ρ* procedure. First, note that when using protocol $\mathcal{H}^{\mathcal{R}}$ with each node having the same ρ, Theorem 8 simplifies to

$$|R_t(v,\bar{\rho})| = C \cdot \min\left(1-\rho, \rho|D(v,\tau)|\right) \quad (4)$$

$$RW_t(\bar{\rho}) = C \cdot \sum_{v \in V} \min(1-\rho, \rho|D(v,\tau)|) \quad (5)$$

Because protocol $\mathcal{H}^{\mathcal{R}}$ has a weak steady state by Corollary 9, maximizing the total remote work is equivalent to maximizing the number of remote queries processed in a single round of the steady state. To distinguish the notation between this case where all nodes have the same ρ from the general case, we use $R(v,\rho)$ instead of $R_t(v,\bar{\rho})$ and $RW(\rho)$ instead of $RW_t(\bar{\rho})$ to signify the special case of identical ρ's. We dropped the time subscript because we are only interested in the steady state.

The proof proceeds in three steps: (1) we establish a range of values for the optimal $\hat{\rho}$, (2) we show $\hat{\rho}$ can only be one of n values within this range where n is the number of nodes, and (3) we then find the optimal $\hat{\rho}$ that maximizes $RW(\rho)$.

For the first step, we ordered the vertices $\{v_1, v_2, \ldots, v_n\}$ such that $|\hat{D}(v_i,\tau)| \leq |\hat{D}(v_{i+1},\tau)|$. We then compute a corresponding sequence $\{d_1, d_2, \ldots, d_n\}$ where $d_i = \frac{1}{|\hat{D}(v_i,\tau)|}$. Notice that the sequence of d_i is non-increasing. We can make the following observation on choosing a particular value of ρ.

Lemma 11. *If $\rho \geq d_i$ for some i, then for all $j \geq i$, node v_j is saturated.*

For example, if we know $\rho \geq d_4$, then we can guarantee nodes v_4 through v_n are saturated. Intuitively, node j has more neighbors than node i if $j > i$. Therefore, if there is enough work to saturate node i, node j is also saturated.

Proof. We want to show that if $j \geq i$, then $\rho|D(v_j,\tau)| \geq 1-\rho$. Since $\rho \geq d_i$, $\rho|D(v_j,\tau)| \geq d_i|D(v_j,\tau)|$ and $1-\rho < 1-d_i$. Therefore, it is sufficient to show that if $j \geq i$, then $d_i|D(v_j,\tau)| \geq 1-d_i$:

$$\begin{aligned} & j \geq i \\ \Rightarrow\quad & |D(v_j,\tau)| \geq |D(v_i,\tau)| \\ \Rightarrow\quad & \tfrac{|D(v_j,\tau)|}{|\hat{D}(v_i,\tau)|} \geq 1 - \tfrac{1}{|\hat{D}(v_i,\tau)|} \\ \Rightarrow\quad & d_i|D(v_j,\tau)| \geq 1-d_i \end{aligned}$$

□

Using the above observation, we complete our first step by bounding the optimal $\hat{\rho}$ between d_1 and d_n, inclusive.

Lemma 12. *The optimal $\hat{\rho}$ is between d_1 and d_n, i.e., $d_1 \geq \hat{\rho} \geq d_n$.*

Proof. For any $\rho > d_1$, Lemma 11 guarantees that all nodes are saturated. Therefore, $RW(\rho) = nC(1-\rho)$. Now consider using $\rho' = d_1$. Lemma 11 still guarantees all nodes are saturated. Thus $RW(\rho' = d_1) = nC(1-d_1)$. Since $\rho > d_1$, $1-\rho < 1-d_1$. Therefore, $RW(d_1) > RW(\rho)$.

For $\rho < d_n$, the converse of Lemma 11 implies no nodes are saturated, hence $RW(\rho) = C \cdot \sum_{v \in V} \rho|D(v,\tau)|$. In comparison to choosing $\rho' = d_n$ where $RW(\rho' = d_n) = C \cdot \sum_{v \in V} d_n \cdot |D(v,\tau)|$, we see $RW(\rho) < RW(d_n)$ because $\rho < d_n$.

Hence, the optimal $\hat{\rho}$ is between d_1 and d_n. □

For the second step of our proof, we refine our search of optimal $\hat{\rho}$ by claiming the optimal $\hat{\rho}$ is in fact d_j for some j. To prove this claim, we show that for any choice of ρ strictly between d_i and d_{i+1} for some i, we can "slide" ρ to one of the two endpoints without reducing the total remote work. More formally,

Lemma 13. *If $d_i > \rho > d_{i+1}$, then either $RW(d_i) \geq RW(\rho)$ or $RW(d_{i+1}) \geq RW(\rho)$.*

Proof. Notice that we can rewrite $RW(\rho)$ as two terms: one term that includes saturated nodes and one term that includes the rest. Using the result from Lemma 11 while knowing $d_i \geq \rho \geq d_{i+1}$, we see that nodes v_{i+1} through v_n are always saturated and nodes v_1 through v_{i-1} are always not saturated. Because $d_i \geq \rho$, $|RW(v_i, \rho)|$ is at most $1 - \rho$, thus it is safe to treat node v_i as a node that is always not saturated. Dividing the nodes into these two categories, we can rewrite $RW(\rho)$ as the sum of work from non-saturated nodes and work from saturated nodes.

$$RW(\rho) = C\left(\sum_{j \leq i} \rho |D(v_j, \tau)| + \sum_{i < j \leq n} (1 - \rho)\right).$$

Now consider the amount of remote work we gain by using ρ rather than using d_i, i.e., $RW(\rho) - RW(d_i)$. Since nodes have the same capacity C, we take out a factor C when computing the gain. We see

$$\begin{aligned}
& \frac{RW(\rho) - RW(d_i)}{C} \\
= \quad & \left(\sum_{j \leq i} \rho |D(v_j, \tau)| + \sum_{i < j \leq n} (1 - \rho)\right) - \left(\sum_{j \leq i} d_i |D(v_j, \tau)| + \sum_{i < j \leq n} (1 - d_i)\right) \\
= \quad & \sum_{j \leq i} (\rho - d_i) |D(v_j, \tau)| + \sum_{i < j \leq n} (d_i - \rho) \\
= \quad & (d_i - \rho)\left(n - i - \sum_{j \leq i} |D(v_j, \tau)|\right) \qquad (6)
\end{aligned}$$

From Eq. (6), using ρ is better than d_i when

$$\begin{aligned}
& RW(\rho) > RW(d_i) \\
\iff \quad & RW(\rho) - RW(d_i) > 0 \\
\iff \quad & n - i > \sum_{j \leq i} |D(v_j, \tau)| \qquad (7)
\end{aligned}$$

Similarly, we get that using ρ is better than using d_{i+1} when

$$\begin{aligned}
& RW(\rho) > RW(d_{i+1}) \\
\iff \quad & n - i < \sum_{j \leq i} |D(v_j, \tau)| \qquad (8)
\end{aligned}$$

In order for both $RW(\rho) > RW(d_i)$ and $RW(\rho) > RW(d_{i+1})$ to hold, Eq (7) and (8) must be true simultaneously, which clearly cannot be the case. Therefore, it is possible to "slide" ρ towards one of the endpoints. $\square$

Combining Lemmas 12 and 13, we complete the final step in Theorem 14.

Theorem 14. *In protocol $\mathcal{H}^{\mathcal{R}}$ with identical ρ, the optimal $\hat{\rho} = d_k$, where k is the smallest integer such that $\sum_{i\leq k} |\hat{D}(v_i, \tau)| \geq n$.*

Proof. We use the insight from the proof of Lemma 13. Suppose we set our initial ρ to d_1. Eq. (7) tells us that using $\rho = d_2$ is better if $n - 1 > |D(v_1, \tau)|$. Similarly, using $\rho = d_3$ is better yet if $n - 2 > |D(v_1, \tau)| + |D(v_2, \tau)|$. We can repeat this step of moving to better d_k until Eq. (7) no longer holds. At that point, we have reached the optimal because Eq. (8) tells us that using d_j for $j > k$ will not improve total remote work.

The formal proof is by contradiction. Suppose d_i is the optimal ρ where $i \neq k$. If $i < k$, then by Eq. (7), using $\rho = d_{i+1}$ results in more remote work, which contradicts d_i being optimal. If $i > k$, then by Eq. (8), using $\rho = d_{i-1}$ yields more work, another contradiction. $\square$

There is a special case for Theorem 14 when $\sum_{i\leq k} |\hat{D}(v_i, \tau)| = n$. In this situation, there are multiple optimal $\hat{\rho}$ for a single round in the steady state. Specifically,

Corollary 15. *If $\sum_{i\leq k} |\hat{D}(v_i, \tau)| = n$ for some k, then for all ρ where $d_k \geq \rho \geq d_{k+1}$, $RW(\rho)$ is optimal.*

Example (d) in Figure 5 illustrates this occurrence of multiple optimal $\hat{\rho}$. The sequence of $\{|\hat{D}(v_i, \tau)|\}_i$ in this case is $\{2, 2, 3, 3\}$. Notice that $|\hat{D}(v_1, \tau)| + |\hat{D}(v_2, \tau)| = 2 + 2 = 4$ which is the number of nodes. By Corollary 15, we can conclude for example (d), any ρ where $\frac{1}{3} \leq \rho \leq \frac{1}{2}$ yields the optimal amount of remote work in a single round of the steady state.

Now that we know how to find the optimal for this special case of identical ρ for each node, a natural question is how much remote work did we sacrifice in restricting to the special case instead of using arbitrary $\bar{\rho}$? To bound this amount of lost remote work, we use the following the theorem.

Theorem 16. *For any connected network $G = (V, E)$ where $|V| = n \geq \tau + 1$, compute the optimal $\hat{\rho}$ using the* Find_Optimal_Single_ρ *procedure. Then in steady state, $RW_t(\hat{\rho}) \geq \frac{\tau}{\tau+1} nC$.*

Proof. Because the graph is connected, for any node $v \in V$, the number of nodes within τ hops is at least $\tau + 1$, i.e., $|\hat{D}(v, \tau)| \geq \tau + 1$. Therefore $d_1 \leq \frac{1}{\tau+1}$. From Lemma 12, we see $RW(\hat{\rho}) \geq RW(d_1) = nC(1 - \frac{1}{\tau+1}) = \frac{\tau}{\tau+1} nC$. $\square$

The immediate consequence of Theorem 16 is that even with the restriction of identical ρ's, nodes in the network are processing at $\frac{\tau}{\tau+1}$ of the maximum capacity. Hence, the fraction of loss due to the restriction is at most $\frac{1}{\tau+1}$. A secondary consequence is that regardless of what kind of network G we use, we can always process remote work at $\frac{\tau}{\tau+1}$ of the capacity. When looking at the proof in more detail, one notice that the bound on the amount of work lost is dependant on the value of d_1, the smallest neighborhood size for a node in the topology. In practice where $\tau \geq 5$, d_1 is at most $\frac{1}{50}$. Thus, the total remote work lost is at most 2%.

To put Theorems 14 and 16 into perspective, we ran a simulation on a 1787-node Gnutella crawl from Saroiu [10] with different choices of ρ value and TTL of 7. We then tallied the total amount of remote work in the network, normalized by the capacity of the nodes (i.e., each node can contribute at most 1 to the total remote work). The results are shown in Figure 6. The x-axis shows the choice of ρ in log scale. The y-axis shows the amount of remote work. As Theorem 14 claimed, there is a single optimal point, $\rho = \frac{1}{75} = 0.0133$ for this topology. Also notice that there is a "reasonably large" stretch of ρ values where the total amount of work is close to the optimal. This stretch corresponds to choosing a larger ρ value than the optimal. As Thereom 16 claims, the amount of remote work lost is not significant for any "reasonable" guesses of ρ, such as $\rho = \frac{1}{50} = 0.02$.

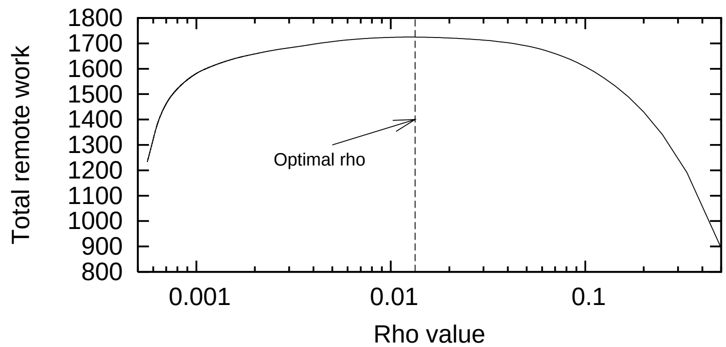


Figure 6: Total remote work as a function of ρ for Gnutella crawl.

8 Different ρ for Each Node

If we have all nodes inject the same number of queries into the network, some nodes will not operate at their maximum capacities. Thus it is possible to achieve more remote work by allowing nodes to inject different amounts of work, i.e., use a different ρ for each node. To illustrate the difference in the amount of remote work, we reuse the examples in Figure 5. In (b), by setting the ρ for the center of the star to 1 and 0 for the other nodes, we can saturate every node and get a total remote work of $6C$. In contrast, the identical-ρ case only yields total remote work of $\frac{7}{2}C$. Similarly, we get $4C$ and $2C$ for examples (c) and (d) respectively by setting the ρ of the nodes with the highest degrees to 1 and 0 for the other nodes. Using identical ρ, we get $\frac{7}{2}C$ and $2C$ respectively for examples (c) and (d).

In this general case where nodes can have different ρ values, there are many possible optimal solutions. In particular, there is one subset of the optimal solutions that corresponds to the *minimum fractional dominating-set* (MFDS) of distance τ for the network topology graph $G = (V, E)$. In MFDS, each node v is assigned a weight w_v where $0 \le w_v \le 1$. The dominating set condition is that for every node v, the sum of the weights from nodes within τ hops of v is at least 1. The goal is to come up with a set of weights w_v that satisfies the dominating condition while minimizing the sum of the weights. The MFDS is a well understood problem. Reducing our problem to the MFDS exposes some underlying structure in finding the optimal ρ_v's and allows us to leverage many existing techniques for solving it. Fortunately, there is a simple mapping from an optimal solution of MFDS to our problem. Specifically,

Theorem 17. *For any optimal solution $\{w_v\}$ to the minimum fractional dominating set of G with distance τ, the solution $\bar{\rho}$ where $\rho_v = w_v$ maximizes the total remote work in the network G.*

Before we prove the above claim, we observe that when all nodes are saturated, maximizing remote work is equivalent to minimizing new-query injection (i.e., MFDS). Therefore we simply need to show that there exists an optimal $\hat{\rho}$ where all nodes are saturated. Intuitively, for any optimal $\hat{\rho}$ where some node v is not saturated, we can "boost" ρ_v until v is saturated without changing the amount of remote work. We now give the details.

First, the minimum fractional dominating-set (MFDS) problem for graph G with distance τ can be rephrased as a linear program. Let w_v be the weight assigned to node v. Then the LP is

$$min : \sum_v w_v \tag{9}$$

$$\sum_{w \in \hat{D}(v,\tau)} w_v \ge 1 \quad \forall\, v \in V \tag{10}$$

From the LP above, notice that when using $\rho_v = w_v$ where $\{w_v\}$ is a solution of MFDS, every node v is saturated because $\sum_{w \in D(v,\tau)} \rho_w = \left(\sum_{w \in \hat{D}(v,\tau)} w_v\right) - w_v \geq 1 - w_v = 1 - \rho_v$. Relying on this observation, we show $\{w_v\}$ is an optimal solution for our problem in two steps: (1) given an optimal solution $\hat{\rho}$ where all nodes are not saturated, we can transform $\hat{\rho}$ into another optimal solution $\hat{\rho}'$ such that the amount of remote work is still the same but all nodes are now saturated; and (2) when nodes are all saturated, minimizing total weight in MFDS is equivalent to maximizing remote work.

To prove the first step, we observe that if we increase ρ_v for a single node v, then the remote work for nodes $w \neq v$ can only increase. Formally,

Lemma 18. *Given $\bar{\rho} = \{\rho_1, \rho_2, \ldots, \rho_n\}$, create $\bar{\rho}' = \{\rho_1', \rho_2', \ldots, \rho_n'\}$ where $\rho_i' = \rho_i + \epsilon$ for some i and $\epsilon > 0$, and $\rho_j' = \rho_j$ for all $j \neq i$. Then $|R(v_j, \bar{\rho}')| \geq |R(v_j, \bar{\rho})|$ for all $v_j \neq v_i$.*

Proof. From Theorem 8, we get $|R(v_j, \bar{\rho}')| = C \cdot \left(\min(1 - \rho_j', \sum_{v_k \in D(v_j,\tau)} \rho_k')\right)$ where the min distinguishes between whether node v_j is saturated or not. Thus for each node $v_j \neq v_i$, we need to check that our claim holds for both cases.

Case 1: v_j is saturated under $\bar{\rho}$. Since the ρ values only increased, v_j is still saturated. Moreover, $p_j = p_j'$. Therefore $|R(v_j, \bar{\rho})| = C \cdot (1 - \rho_j) = C \cdot (1 - \rho_j') = |R(v_j, \bar{\rho}')|$.

Case 2: v_j is not saturated under $\bar{\rho}$. Then $|R(v_j, \bar{\rho})| < C \cdot (1 - \rho_j) = C \cdot (1 - \rho_j')$ and $|R(v_j, \bar{\rho})| = C \cdot \sum_{v_k \in D(v_j,\tau)} \rho_k \leq C \cdot \sum_{v_k \in D(v_j,\tau)} \rho_k'$. Since $|R(v_j, \bar{\rho})|$ is less than both $C \cdot (1 - \rho_j')$ and $C \cdot \sum_{v_k \in D(v_j,\tau)} \rho_k'$, we get $|R(v_j, \bar{\rho})| \leq C \cdot \left(\min(1 - \rho_j', \sum_{v_k \in D(v_j,\tau)} \rho_k')\right) = |R(v_j, \bar{\rho}')|$. ☐

Note that the claim does not hold for node v_i because if v_i is already saturated under $\bar{\rho}$, then increasing ρ_i will reduce the remote work at node v_i under $\bar{\rho}'$. Intuitively, using Lemma 18, we can "boost" the ρ_v values of non-saturated nodes, one by one, in the optimal solution while maintaining the same number of total remote queries processed. Specifically,

Lemma 19. *For every optimal $\hat{\rho} = \{\rho_1, \rho_2, \ldots, \rho_n\}$ where some nodes are not saturated, there exists a corresponding optimal $\hat{\rho}' = \{\rho_1', \rho_2', \ldots, \rho_n'\}$ such that $RW(\hat{\rho}) = RW(\hat{\rho}')$ and all nodes are saturated using $\hat{\rho}'$.*

Proof. Given $\hat{\rho} = \{\rho_1, \rho_2, \ldots, \rho_n\}$, suppose node $v_i \in V$ is not saturated, i.e., $R(v_i, \hat{\rho}) < 1 - \rho_i$. Construct $\hat{\rho}' = \{\rho_1, \ldots, \rho_{i-1}, \rho_i', \rho_{i+1}, \ldots, \rho_n\}$ where $\rho_i' = 1 - R(v_i, \hat{\rho})$.

By construction, $\hat{\rho}_i' > \hat{\rho}_i$ and does not reduce the remote work at node v_i. By Lemma 18, using $\hat{\rho}_i'$ does not reduce the remote work for all nodes $v_j \neq v_i$. Therefore, $RW(\hat{\rho}') \geq RW(\hat{\rho})$. Because $\hat{\rho}$ is optimal, $RW(\hat{\rho}') \leq RW(\hat{\rho})$. Hence, $RW(\hat{\rho}') = RW(\hat{\rho})$.

Using $\hat{\rho}'$ results in at least one more saturated node than $\hat{\rho}$. By repeating the above step of boosting one node's ρ, we can construct an optimal where all nodes are saturated. (Note that multiple boosting steps cannot be applied simultaneously. One must apply each boost in sequence and identify a new node to boost each time.) ☐

Lemma 19 completes our first step for showing that there exists an optimal $\hat{\rho}$ where all the nodes are saturated. We now show our second step, the proof of Theorem 17, where minimizing total weight in MFDS is the same as maximizing remote work.

Proof. From Lemma 19, we can assume every node v is saturated, which by Theorem 8 occurs precisely when the sum of the work injected by nodes in $\hat{D}(v, \tau)$ is at least C. Scaling down by a factor of C yields the constraint for MFDS in Eq. 10.

Since every node in the network is saturated, the total work in the entire network (i.e., the sum of local and remote work) is equal to the number of nodes in G, a constant. Thus maximizing remote work is equivalent to minimizing local work. Since local work is $\sum_i \rho_i$, we get the same objective function as MFDS in Eq. 9. Therefore, any optimal solution of the MFDS maximizes the total remote work in G. $\square$

Although using different ρ's leads to more remote work, note that we are setting ρ to 0 for a large number of nodes, which means these nodes cannot inject any queries. In practice, a node that cannot inject any queries is not useful. Therefore a combination of using a small fixed ρ (e.g., using d_n from the previous section) to guarantee some fairness while allocating the remaining capacity through the dominating set is more practical.

9 Open Problems

We now outline two open problems that are practical variations of the maximizing remote work problem we studied in this paper.

9.1 Distributed Algorithm

In Sections 7 and 8, we described centralized solutions for finding the optimal ρ for each node that maximizes the total remote work in the network. Our solutions require knowing the entire network topology in advance. However in a P2P environment, with nodes constantly joining and leaving, it is impractical for any node to gather the entire network topology information. Even if we could efficiently gather such information, the rapidly changing topology will quickly render a solution based on the current topology obsolete and sub-optimal. Nevertheless, the results about the centralized solutions are important because they form the basis of comparison for distributed solutions.

For the instance of using a different ρ for each node, distributed solutions are possible by adapting fractional dominating set algorithms [2], [6]. However, these algorithms have long running times for our problem, cannot handle different capacities at each node, and must be re-run each time as the network topology changes. Here, we propose a simple heuristic for estimating how many new queries each node should inject (i.e., the value of $\rho_v C$ for each node v) in a distributed fashion. Figure 7 outlines the steps in our distributed approach. Every node only makes local decisions. When a node does not have enough queries to saturate its processing capacity, it tells all of its neighbors to inject one more local query per round. If a node has too much remote work, it tells all the nodes that have sent remote work to it to inject one less local query per round. We have performed some initial simulations to compare our heuristic against the optimal solution. The heuristic performs very well when the capacity C, in number of queries, is large compared to the number of nodes within τ hops.

The randomization for $inc(p)$ and $dec(p)$ is necessary to avoid oscillation and to stabilize the system. However, the resulting stable setting may not be optimal. Hence, a better solution is needed. However, note that the proposed heuristic is estimating the number of new queries ρC rather than the fraction of capacity ρ as in the fraction dominating set approach. Thus this heuristic does not assume all the nodes have the same capacity C.

9.2 Nodes with Different Capacities

In reality, super-nodes may have different processing capacities. The results from the previous sections no longer hold because we cannot determine, independent of the network topology, when a node is saturated. Recall that if nodes have the same capacity, then Lemma 7 guarantees that a node v is saturated when v's

Distributed ρC Estimation

For every 2τ rounds (say at time t), each node $v \in V$ does the following:

1. If $|W_t(v, \rho_v)| < C$ (i.e., not enough remote work),
2. broadcast an $inc(1 - \frac{|W_t(v,\rho_v)|}{C})$ message with TTL τ.
3. If $|W_t(v, \rho_v)| > C$ (i.e., too much remote work),
4. for every node w such that $\exists(w, ttl, mid) \in W_t(v, \rho_v)$
5. send a $dec(\frac{|W_t(v,\rho_v)|}{C} - 1)$ message to node w.

Upon receiving an $inc(p)$ or $dec(p)$ message, each node adjusts its ρC by 1 with probability p.

Figure 7: An informal description of a distributed ρC estimation heuristic.

neighbors are injecting more queries than v's capacity. However, when nodes have different capacities, there is a simple counterexample.

Consider nodes u, x, and v connected in a line in that order. Now assign capacity $2C$ to nodes u and v and capacity C to x. Since all the work from u must travel through x to reach v, the amount of remote work at v is limited by the capacity at x. Even if node u is injecting $2C$ queries, at most C of them will reach v each round, which invalidates Lemma 7 for the case of different capacities. In this particular example, the extra capacities at nodes u and v are irrelevant.

Even for the simple case where only one node x has more capacity than the rest, the solution is non-obvious and topology dependent. For example, if x is in an area of the network where nodes are under-saturated, then it should use its extra capacity to inject more queries. On the other hand, if x is in an area where nodes are already saturated, then the extra capacity should only be used to increase the amount of remote work at node x.

Our current approach is an incremental heuristic that combines multiple optimal solutions. The basic idea is as follows: Suppose nodes have one of two possible capacities C_1 and C_2 where $C_1 < C_2$. Then our heuristic is to find the optimal $\bar{\rho}$ setting for the entire network assuming all the nodes have capacity C_1. We then create a subgraph of the original network that includes only nodes with capacity C_2. Note that the subgraph may be disconnected. We then compute another optimal $\bar{\rho}'$ setting on the subgraph assuming all the nodes have the capacity $C_2 - C_1$. For nodes with capacity C_1, their corresponding ρ' value is 0. To get the final solution, we let each node inject $\rho_v C_1 + \rho'_v(C_2 - C_1)$ queries.

10 Concluding Remarks

This paper uses a simple model to study remote work in a flooding-based peer-to-peer network. In particular, we showed

1. For any setting $\bar{\rho}$, protocol $\mathcal{H}^{\mathcal{R}}$ processes the most remote work.
2. Under protocol $\mathcal{H}^{\mathcal{R}}$ with all nodes using the same ρ, if we order the nodes $\{v_1, \ldots, v_k\}$ where $|\hat{D}(v_i, \tau)| \leq |\hat{D}(v_{i+1}, \tau)|$, then the optimal $\hat{\rho} = \frac{1}{|\hat{D}(v_k,\tau)|}$ where k is the smallest integer such that $\sum_{i=1}^{k} |\hat{D}(v_i, \tau)| \geq n$.
3. When nodes use different ρ, any optimal solution to the minimum fractional dominating-set of the network graph G is an optimal $\hat{\rho}$ solution.

We believe that our results can serve as a benchmark for more complex systems. For example, the proposed heuristic load management scheme of Section 9.1 can be compared against a system where $\bar{\rho}$ is selected using our optimal and centralized solutions. In addition, our solutions can form the basis for heuristics, as illustrated in Section 9.2.

Acknowledgment We thank Kamesh Munagala for valuable discussions on approximation algorithms for fractional bin-packing.

References

[1] N. Daswani and H. Garcia-Molina. Query-flood DoS attacks in Gnutella. In *ACM Conference on Computer and Communications Security*, Washington, DC, November 2002.

[2] N. Garg and J. Könemann. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. In *39th Annual Symposium on Foundations of Computer Science*, pages 300–309, Palo Alto, California, November 1998.

[3] The Gnutella Developer Forum (GDF). Database of vendor codes. http://groups.yahoo.com/group/the_gdf/.

[4] Gnutella. Website http://gnutella.wego.com.

[5] Concurrent Gnutella Hosts. http://www.limewire.com/.

[6] F. Kuhn and R. Wattenhofer. Constant-time distributed dominating set approximation. In 22nd *ACM Symposium on Principles of Distributed Computing*, Boston, MA, July 2003.

[7] 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 annual ACM International Conference on Supercomputing (ICS)*, 2002.

[8] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. In *Proceedings of ACM SIGCOMM*, pages 149–160, San Diego, August 2001.

[9] 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.

[10] S. Saroiu, P. K. Gummadi, and S. D. Gribble. Measuring and analyzing the characteristics of Napster and Gnutella hosts. In *Multimedia Computing and Networking (MMCN)*, San Jose, CA, January 2002.

[11] 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 ACM SIGCOMM*, pages 160–177, San Diego, August 2001.

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

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