Mitigating Information Asymmetries to achieve Efficient Peer-to-Peer Queries

Jiang Guo, Baochun Li
Department of Electrical and Computer Engineering
University of Toronto
{jguo,bli}@eecg.toronto.edu

Abstract

Querying for a particular data item is perhaps the most important feature to be supported by peer-to-peer network infrastructures, and receives the most research attention in recent literature. Most existing work follows the line of designing decentralized algorithms to maximize the performance of peer-to-peer queries. These algorithms often have specific rules that peer nodes should adhere to (e.g., placement of data items on particular nodes), and thus assumes that peers are strictly cooperative. However, in realistic peer-to-peer networks, selfish and greedy peer nodes are the norm, and query strategies degenerate to random or flooding based searches. In this paper, we explore the design space with respect to query efficiency in selfish peer-to-peer networks where nodes have asymmetric information, *and apply the* signaling *mechanism from microeconomics to facilitate the sharing of private information and thus improve search efficiency. We extensively simulate the signaling mechanism in the context of other alternative solutions in selfish networks, and show encouraging results with respect to improving query performance.*

1 Introduction

Querying for a particular data item is perhaps the most important feature to be supported by peer-to-peer network infrastructures, and receives the most research attention in recent literature. Most existing work seeks to design decentralized algorithms to maximize the performance of peer-to-peer queries. These algorithms often have specific rules that peer nodes should adhere to. Examples of such rules include all research proposals in the area of *structured* peer-to-peer networks, which specify the mandatory placement of data items on particular nodes (e.g., the Chord protocol [10]), and thus assist to achieve querying performance in the order of $\log N$, N being the size of the peer-to-peer network. Obviously, these proposals assume that peers are strictly *cooperative* when it comes to implementing these rules. However, in realistic peer-to-peer networks, selfish and greedy peer nodes are the norm, and query strategies unfortunately degenerate to random or flooding based searches (e.g., Gnutella). In such scenarios (often referred to as *unstructured* peer-to-peer networks), either the search performance is not satisfactory ($O(N)$) when doing random searches, or the message exchange overhead is high, when performing flooding-based searches.

Various previous work [1, 4, 6, 11] have been attempting to address such problem of querying performance in unstructured peer-to-peer networks. As one example, Cohen *et al.* [4] resort to the approach of requiring peer nodes to *collaboratively replicate* the actual data items in demand. We observe that all previous proposals have resorted to the introduction of a certain degree of *structure* or *discipline* (e.g., required replications) to an unstructured peer-to-peer network, and with this measure they have been successful in achieving better querying performances. Furthermore, we emphasize the following key observation: Introducing a certain degree of structure or discipline in unstructured networks has imposed the unwarranted assumption of *cooperation*, which is against the observation that leads to the study of such networks in the first place: *selfishness of nodes*. For one example, selfish peer nodes may not even be willing to share its own *private information* to others (the case of *asymmetric information*), not to mention the cooperation required to implement protocols that introduce structure (e.g., replications).

In this paper, we explore the following questions: What may occur if peer nodes do not share their private information, such as shortcuts to existing data items? What may be possibly proposed to incentivize selfish nodes to share their private information for the common good? Ultimately, what may be possibly proposed to improve query efficiency in selfish peer-to-peer networks where nodes hold asymmetric information? In light of all these questions, we propose to apply the *signaling* mechanism from microeconomics to facilitate the sharing of private information and thus improve search efficiency and avoid the phenomenon of adverse selection. We extensively simulate the signaling mechanism

in the context of other alternative solutions in selfish networks, and show encouraging results with respect to improving query performance.

The remainder of the paper is organized as follows. We formulate the problem in Sec. 2. In Sec. 3 we begin by discussing the case of asymmetrical information, and propose mechanisms to signal those private information, which can significantly improve the system equilibrium. In Sec. 4 we propose a set of distributed algorithms and present simulation results. Sec. 5 concludes the paper.

2 Problem Formulation

In this paper, we consider an *unstructured peer-to-peer network* that consists of N selfish nodes. Each node can create shortcuts (*i.e.*, pointers) to any data items on any nodes. In addition, each node also has a set of *neighbors*, which generally includes all the other nodes in the network that this node is aware of. Since on each node u the table to contain shortcuts is of limited size which is smaller than the number of items in the network, when node u wishes to access an item that is not in its own shortcut table, node u has to turn to its neighbors to forward the query until a node that owns such item is found or certain maximum query limits are imposed. We illustrate the nodes and their roles in the peer-to-peer network in Fig. 1.

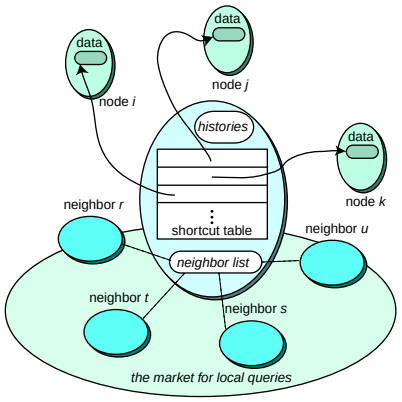

Figure 1. Nodes in a peer-to-peer network.

Provided that nodes are selfish and always seek to maximize their own gains, it is natural that nodes have to be rewarded in some way for providing the service of processing incoming queries. For convenience of our analysis, we model such queries in a *market*. At each query hop, the node that initiates the query and its neighbors constitute the market. In such market, the querying node, referred to as a *consumer*, is interested in the *commodity*, which is the information on the item's location. As a consumer, the querying node can purchase the commodity from its neighbors, also referred to as *producers*.

The nodes that initiate the queries wish to access the item as quickly and efficiently as possible — they benefit or *gain* from such successful queries. We introduce θ to quantitatively model such gains, and to reflect the quality of the purchased commodity. A commodity is said to have a quality of θ, if the consumer receives a gain of θ when the consumer purchases such a commodity. A higher θ implies that the producer providing such a commodity is more likely to find the item. Without loss of generality and for convenience of our analysis, the producer that provides a commodity of θ gain is considered to have a *type* of θ (using microeconomics terms). We let $[\underline{\theta}, \overline{\theta}] \subset \mathbb{R}$ denote the set of possible quality levels, where $0 \leq \underline{\theta} < \overline{\theta} < \infty$. $\overline{\theta}$ implies that the producer is most likely to locate the item, *i.e.*, it has the item in its possession or a shortcut to the item in its shortcut table. On the contrary, $\underline{\theta}$ implies that the producer is most unlikely to find the item, *i.e.*, it has no idea with respect to the whereabouts of the item. When selecting a neighbor to forward the query, it is clear that the neighbor of the highest type is always preferred.

The proportion of producers with θ or less is given by the distribution function $F(\theta)$. For simplicity, we assume that $F(\cdot)$ is nondegenerate and has an associated density function $f(\cdot)$, with $f(\theta) > 0$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$. To produce a commodity of quality θ (to obtain the information on the position of the item, *e.g.*, entries in the shortcut table with a limited size), a producer incurs certain costs, denote as $r(\theta)$. We make the following assumptions with respect to such costs:

1. $r(\theta) \leq \theta$, for all $\theta \in [\underline{\theta}, \overline{\theta}]$;
2. $r(\cdot)$ is a strictly increasing function.

The first assumption implies that every producer has a chance to accept queries, so that we disregard the case when $r(\theta)$ is larger than θ. The second assumption implies that the producers with higher types are more likely to cost more to produce the commodity, *i.e.*, to obtain the information of the item. This assumption is mainly for the simplicity of our theoretical analysis. We relax such an assumption in our simulation experiments in Sec. 4.2.

Consider a producer v of type θ, θ is proprietary to v itself, and not known by other nodes, including its neighbors. We refer to such information as *private information*. Such a scenario where nodes keep private information from other nodes is economically known as a system with *asymmetric information*. The opposite case is the system with *symmetric information*. Asymmetric information significantly impairs query performance in peer-to-peer networks. For example, suppose a consumer u chooses to purchase the commodity at a price p from a set of producers with different types. In the case of symmetric information, the consumer u chooses the producer with the highest type $\overline{\theta}$ and receives the profit of $\pi_u = \overline{\theta} - p$. However, in the case of asymmetric information, the consumer can not tell the producers of

high types apart from the producers of low types, *i.e.*, the consumer may not be able to choose the producer with $\overline{\theta}$.

In this paper, we are interested in exploring the effects of asymmetric information, as well as different alternatives of mitigating asymmetries in order to improve peer-to-peer queries.

3 Mitigating Information Asymmetries

For comparison purposes, we first consider the case of symmetric information, the ideal case where the types of the neighbors are *publicly observable*. In this case, since the consumer u knows the quality of all commodities, it is clear that u will simply choose the commodity of the highest quality. By choosing the neighbor that has the highest type θ, the consumer u will maximize its payoff.

To study the price p^* at the equilibrium, we study the market where multiple consumers compete for one commodity by setting different prices, and the one with the highest price obtains the commodity. Following the *Bertrand model* [3], this is a simple case of an oligopolistic market, where we have $p^* = \overline{\theta}$ at equilibrium. The consumer u earns zero profit due to the competition from other consumers. The producer of the highest type $\overline{\theta}$ wins the offer and obtains a welfare of $\overline{\theta} - r(\overline{\theta})$. It is clear that upon this query the market is efficient. However, such an ideal case is unrealistic in unstructured peer-to-peer networks, where such type information is not publicly observable, belonging to the case of asymmetric information.

3.1 Information asymmetries: microeconomics

We proceed to study the equilibrium in the case of asymmetric information and the welfare at such equilibrium.

The problem is modeled as a dynamic game $\mathcal{G}$. The steps of the game are as follows: (1) the consumer u claims a price p; (2) each producer chooses to accept it or not and replies its choice to u; and (3) u randomly chooses one of the set of neighbors who accept the price.

Due to asymmetric information, the consumer u does not have any knowledge about its producers except $\underline{\theta}$ and $\overline{\theta}$. We note that when the types of producers are not observable, the price p must be independent of such types.

From the perspective of a producer with type θ, its strategy is straightforward: it chooses to serve u only if $r(\theta)$ is equal to or smaller than p.

Let the inverse function of $x = r(\theta)$ be $\theta = g(x)$, the average commodity quality that the consumer u receives when price is p will be $D(p) = \int_{\underline{\theta}}^{g(p)} xf(x)dx$, where $f(\cdot)$ is the density function of producers' type.

Particularly, for any $p \in [r(\overline{\theta}), \overline{\theta}]$, all producers will choose to accept the price. For simplicity, we let $E[\theta]$ represent the expected commodity quality that the consumer u receives when all the producers are willing to accept the price, *i.e.*, $E[\theta] = \int_{\underline{\theta}}^{\overline{\theta}} xf(x)dx$.

To examine the consumer's strategy, we introduce the notion of a *competitive equilibrium* presented in Definition 3.1 1.

Definition 1: When producer types are unobservable, a *competitive equilibrium* is an operating point, where $p^* = D(p^*)$.

We call this a microeconomic approach, in which case the consumer u will choose to be at the competitive equilibrium.

Proposition 1: There exists at least one competitive equilibrium, where $p^* = D(p^*)$.

Proof: Let function $\phi(p)$ represent $D(p) - p$. Since $D(p)$ is continuous on $p \in [r(\underline{\theta}), \overline{\theta}]$, $\phi(p)$ is continuous on the same range as well. When $p = r(\underline{\theta})$, $D(p) = \underline{\theta}$; hence, $\phi(r(\underline{\theta})) = \underline{\theta} - r(\underline{\theta})$. When $p = \overline{\theta}$, $D(p) = E[\theta]$ and $\phi(\overline{\theta}) = E[\theta] - \overline{\theta}$. With assumption 1, we have $r(\theta) \leq \theta, \forall\theta \in [\underline{\theta}, \overline{\theta}]$. Therefore, we have $\phi(r(\underline{\theta})) \geq 0$. On the other hand, it is not difficult to find out $E[\theta] \leq \overline{\theta}$, such that we have $\phi(\overline{\theta}) \leq 0$. According to the *intermediate value theorem* [5], there exists at least one p^* that satisfy $\phi(p^*) = 0$, *i.e.*, $p^* = D(p^*)$. □

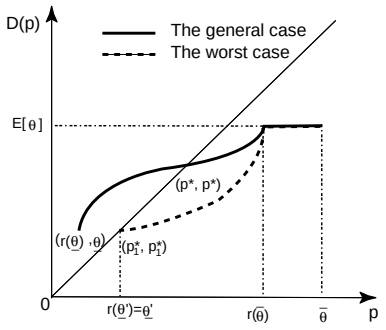

Figure 2. General competitive equilibrium and the worst case

Fig. 2 illustrates how Definition 1 helps to determine competitive equilibria. The solid curve is the general case of function $D(p)$, which has a competitive equilibrium p^*. Furthermore, from Fig. 2, we have two observations:

1. Competitive equilibria are actually those intersecting points between the curve $D(p)$ and the line of $D(p) = p$;

2. The competitive equilibrium price p^* is upper bounded by $E[\theta]$ and lower bounded by $\underline{\theta}$.

The first observation is straightforward, according to the definition of competitive equilibrium. From the second observation, we notice that the competitive equilibrium may not be efficient and may be much smaller than $E[\theta]$. The problem is that to get the producers of the highest type to

accept the query, we need p to be at least $r(\overline{\theta})$. However, the consumer u can not reach even at this price because doing so will bring u a negative payoff of $E[\theta] - r(\overline{\theta})$. Therefore, the consumer will choose to lower the price p until the competitive equilibrium is encountered. In microeconomics term, such phenomenon is known as *adverse selection*. In the present context, adverse selection arises when only the commodities of lower quality are chosen, *i.e.*, only the neighbors that have worse path to the item are willing to accept the query. The worst case occurs, which is depicted as the dotted curve in Fig. 2, when we have $r(\underline{\theta}) = \underline{\theta}$ and $r(\theta) < \theta$ for all other θ. The equilibrium price is $p^* = \underline{\theta}$, which is the lower bound.

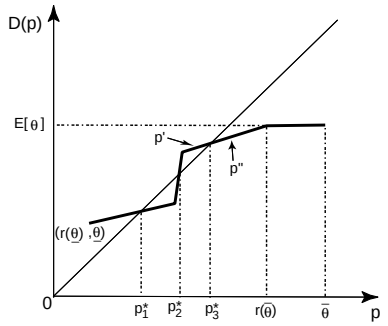

Figure 3. The case of multiple competitive equilibria

Depending on the distribution of neighbors/producers of different types $F(\theta)$, multiple competitive equilibria may exist, *e.g.*, Fig. 3 depicts a case in which there are three equilibria. Unfortunately, the lowest competitive equilibrium is always chosen by such microeconomic approach. In microeconomics, this is called *coordination failure*. The reason is straightforward: the price p is too low because the consumer u expects that the average quality of the commodities is low and, at the same time, only low type producers accept the query precisely because the price p is low.

To summarize, based on the concept of competitive equilibrium, the microeconomic approach is inefficient due to *adverse selection* and *coordination failure*. Because of adverse selection, the average quality of commodity that the consumer u receives can not reach the upper bound $E[\theta]$, while coordination failure causes the consumer u to choose to stay at the lowest competitive equilibrium.

3.2 Information asymmetries: games

We propose an alternative approach based on game theory to address the problem of coordination failure. In such a game-theoretic model, the consumer u could change the price p and choose not to enter an equilibrium, if it observes that deviating from such equilibrium can yield higher payoff. In other words, in the game-theoretic model, consumer u is more sophisticated: If the price is too low, the consumer u will find it in its interest to offer a higher price and attract better neighbors until the highest-price competitive outcome is obtained.

Furthermore, we will show that the consumer u chooses to stay at the highest competitive equilibrium, which is actually the unique subgame perfect Nash equilibrium defined by Proposition 2.

Proposition 2: Let p^* be the highest competitive equilibrium. If there is an $\epsilon > 0$ such that $D(p') > p'$ for all $p' \in (p^* - \epsilon, p^*)$, then p^* is the unique pure strategy subgame perfect Nash equilibrium of the game-theoretic model.

Proof: We observe that Proposition 2 is exactly the first case of Proposition 13.B.1 in [7]. Refer to [7] for a detailed proof. □

For example, in Fig. 3, consider the equilibria p_1^* and p_2^*, which are obviously dominated by the equilibrium p_3^*. In equilibrium p_3^*, for any $p'' > p_3^*$, the consumer u will choose p_3^* due to adverse selection, *i.e.*, p'' is dominated by p_3^*. On the other hand, for any $p' \in (p_2^*, p_3^*)$, the consumer u earns positive profit on p'. However, due to the competition from other consumers, the consumer u will finally choose p_3^* instead and earn zero profit. Therefore, the consumer u will arrive at the highest equilibrium p_3^* and will no longer deviate.

The game-theoretic model solves the issue of coordination failure and achieves the highest competitive equilibrium. However, the game-theoretic model still suffers from adverse selection that is originated from asymmetric information. Consequently, the outcome of the game-theoretic model is also upper bounded by $E[\theta]$.

3.3 Mitigating Information asymmetries: signaling

Signaling [2, 9], as a branch of microeconomics, has been proposed to solve the problem of asymmetric information. In this paper, we apply the *signaling* mechanism to accomplish this objective. The basic idea is that the producers (neighbors) of high types may have actions they can take to distinguish themselves from their low type counterparts.

We model the problem as a *signaling game*, also a dynamic game of asymmetric information. As an extreme example to simply our analysis, we ask the producers to *signal their types by actually finding the item*. Obviously, this approach results in much overhead, since in some cases a number of producers would like to perform the query. In Sec. 4, we introduce a refined approach whose overhead is more moderate. We adapt the model discussed above as follows. The steps of the dynamic game is as follows: (1) the consumer u claims a price p and asks the producers — all its neighbors — to find the item; (2) each neighbor (producer)

chooses whether or not it should *exactly* locate the item and replies the result to the consumer u; and (3) u randomly choose one of the neighbors which has reported positive results to process the query.

In the analysis that follows, we examine the possible equilibrium of the signaling model and the welfare of each producer and consumer. Compared with the model discussed in Sec. 3.1, the significant extension of our signaling model is that the producers can infer their types by executing the signaling action. The cost incurred by such an action for a type θ producer is given by the function $c(\theta)$, which is assumed to be lower for high type producers, *e.g.*, $c(\overline{\theta}) < c(\underline{\theta})$. Since in most cases $r(\theta) \ll c(\theta)$, to keep things simple, we concentrate on the special case in which $r(\theta) = 0, \forall\theta \in [\underline{\theta}, \overline{\theta}]$.

From the perspective of a producer v with type θ, its strategy is as follows:

- v chooses to take the signaling action, if $c(\theta) \leq p$;
- otherwise, v chooses not to take the signaling action:
  - if there exists another producer that takes signaling action, v is considered by the consumer u as low type producer; therefore, v receives zero profit.
  - if none of producers take signal action, the model degenerates to the microeconomic model discussed in Sec. 3.1.

For convenience of analysis, We further assume that $c(\overline{\theta}) < \overline{\theta} < c(\underline{\theta})$, so that only part of producers take the signaling action. Let the inverse function of $x = c(\theta)$ be $\theta = h(x)$, the average commodity quality that the consumer u receives when price is p will be $D'(p) = \int_{h(p)}^{\overline{\theta}} xf(x)dx$, where $p \in [c(\overline{\theta}), c(\underline{\theta})]$.

Similar to Sec. 3.1, we have the following definition and proposition.

Definition 2: When producers are eligible to take signaling actions, a *signaling equilibrium* is an operating point, where $p^* = D'(p^*)$.

Proposition 3: If there exists at least one producer that takes the signaling action, there exists at least one signaling equilibrium, where $p^* = D'(p^*)$.

Due to space constraints, we omit the proof that is similar to the proof of Proposition 1.

Since the consumer u should earn at least zero profit, p^* can not be larger than $\overline{\theta}$. Such an upper bound can be achieved when $c(\overline{\theta}) = \overline{\theta}$. Therefore, p^* is upper bounded by $\overline{\theta}$.

We then discuss two extreme cases: (1) all producers take the signaling action; (2) none of producers take the action. In the former case, $c(\underline{\theta}) \leq p$, *i.e.*, the signaling action is so inexpensive that each producer takes the action attempting to distinguish itself from other producers; on the contrary, in the latter case, $c(\overline{\theta}) > p$, meaning that the signaling action is too costly for any producer to take. However, in both cases, since we assume $r(\theta) = 0$, all producers accept the offer p. Furthermore, since the consumer u can not distinguish high type producers from low type ones, the expected commodity quality that the consumer u receives is $E[\theta]$ in both cases. If we consider both cases as extreme cases of the signaling game, the signaling equilibrium price p^* is lower bounded by $E[\theta]$.

In most cases when only part of producers takes the signaling action, the signaling approach is efficient: the signaling approach can distinguish the producers of higher types and consequently improve the system query efficiency. However, signaling also incurs costs. For example, the proposed signaling approach may be costly in term of overhead. Consider a case when most neighbors have high types and would like to distinguish themselves, most neighbors would choose to complete the query; therefore, this approach becomes a flooding style query, which results in significant message overhead.

4 Algorithm Design and Performance Evaluation

4.1 HU algorithm: type estimation

In previous sections, in the case of a query, we assume that each producer has a type θ, which reflects the extent to which the neighbor knows the position of the item being queried. Obviously, such type does not exist in reality. For the purpose of providing type information, we introduce HU algorithm, which is required before any microeconomics-based approaches may be implemented.

In HU algorithm, each node is required to keep a history of the queries. The history contains pair values of the following information: (1) the identifier of an item; (2) the timestamp of the last query to this item. When a node has just performed a query on an item, or forwarded a query on such item, the node updates the timestamp of this item. The intuition behind such mechanism is that a node that has recently performed or forwarded a query on an item is more likely to have an advantage over another node that has not recently encountered a query on such an item. Since the size of the history is limited, when a new entry is about to be inserted and the history is full, the history entry with the lowest timestamp is evicted.

Given the history of queries, we can obtain the type θ from the timestamp t of the query. Consider an incoming query on item i, the producer v checks its local storage as well as its history of queries. In the case that item i has an entry in the query history with timestamp t, a larger t means higher probability to achieve a successful query, indicating

a higher type θ. Formally, we can assume that producer v associates θ with t through a mapping function $\theta = \sigma(t)$, which increases on t. In the case that item i is possessed by v, we can consider t to be the current time t_{now}. Another extreme case occurs when item i is in neither local storage nor the query history, we let t be zero in this case.

Similarly, we can estimate $r(\theta)$ and $c(\theta)$. The production cost $r(\theta)$ has two parts: (1) the occupation cost of the entry in the query history; (2) the cost to update the entry. We model the production cost $r(\theta)$ by the function $r(\theta) = \tau(L_v, x, \theta)$, where L_v is the size of the query history of v, and x is the number of times that item i's history entry has been updated during last time period T. A larger history size leads to lower $r(\theta)$, while a higher updating frequency means higher $r(\theta)$.

In order to estimate $c(\theta)$, we require that the current producer needs to know the number of hops h that current query has covered so far. This can easily be done by allocating an entry in the message header of such a query message and asking intermediate producers to increment such entry. We model the signaling cost $c(\theta)$ by the function $c(\theta) = \chi(h, \theta)$. Intuitively, $c(\theta)$ becomes lower when h increases.

4.2 Performance Evaluation

We have conducted simulation-based experiments using a packet-level, event-based C++ simulator to evaluate the effects of asymmetric information and to reveal the strengths of the signaling algorithm (SG) (Sec. 3.3), in the context of three approaches: (1) The ideal case of symmetric information (SI); (2) the microeconomic approach illustrated in Sec. 3.1 (ME); and (3) the game-theoretic approach outlined in Sec. 3.2 (GT).

SI is the ideal case that can always find the producer that has the highest type. In the other algorithms, transactions occur among nodes during the query. The difference is that in ME the consumer always adheres to the lowest competitive equilibrium while in GT the highest competitive equilibrium is chosen. In SG, the producers are offered an opportunity to complete specific actions in order to signal their types.

The signaling approach proposed in Sec. 3 is effective to enable high type producers to distinguish themselves. However, such approach may lead to significant message overhead, especially in the case that most producers are of high type. Instead, we introduce a slightly different version of such a signaling method. The consumer incrementally increases the price p until one or more producers accept to process the query or $p_{\max}$ is reached (obviously $p_{\max} \leq \overline{\theta}$); the consumer randomly chooses one of them and pay the price p until the query completes. Such refined approach does not qualitatively change the purpose of the signaling method and, meanwhile, significantly reduces the message overhead brought by signaling.

Initially, each node is given a set of peer nodes as neighbors, which consists of an initial topology. To focus on the effect of shortcuts, we fix the neighbor set during the simulation. We have performed most of our simulations in a ring-like initial topology. We have also performed the simulation in the cases of other initial topologies such as (1) two-dimensional grids; (2) random graphs with Zipf-like node degree distributions; and (3) random graph with Gnutella-like node degree distributions, and have obtained qualitatively similar results. We omit the figures of these settings due to the space constraints.

The network consists of 1000 homogeneous nodes and 100 items. Each node has a shortcut table of 20 and a history [1]of size 40, *i.e.*, $L = 40$. All items have uniform popularity. The network generates queries using a Poisson process with an average query rate λ of 100 queries per time unit. The life time of each item follows the exponential distribution with the average life time of $\mu = 100$ time units. Each simulation runs 1000 time units which is sufficient for each algorithm to reach a stable state.

We assume that initially each item only has one copy in the network, and each consumer never replicates the item that has been queried and supposedly retrieved. The querying process progresses until the desired item is located or the maximum hops limit of 40 is reached.

We list the estimation functions that we use in the simulations as follows [2].

$$\theta = \frac{\max(t - t_0, 0) * 1000}{T} \quad (1)$$

$$r(\theta) = \frac{\min(x + 1, 20) * \theta}{20 * L} \quad (2)$$

$$c(\theta) = \frac{(40 - h) * 500 * T}{40 * \max(1, t - t_0)} \quad (3)$$

where t is the timestamp of the history entry, t_0 is the beginning moment of every time period of T ($T = 50$).

4.2.1 Simulation Results: Query Efficiency

With respect to the efficiency of queries, we compare all four algorithms in both the average query resolution hops and the query success ratios.

Fig. 4(a) shows that none of the algorithms has succeeded to resolve all the queries[3]. This is reasonable since the network is *strictly* unstructured with a query hop limit

[1]In case when an item is not in the history, it means the node has no idea on the whereabout of such item, *i.e.*, it has the lowest type.

[2]We believe that substituting the constants in the functions with different values does not qualitatively effect our conclusions

[3]Since the type information we use here is not accurate and just an approximation, the SI algorithm in our simulation cannot always resolve queries.

of 40. The uniformly distributed item popularity and the approximation of type information also contribute to such results. SI always outperforms all the other alternatives, for the reason that a node in SI can always select the neighbor with the best knowledge of the desired item. SG is the second best, which solves about 40% compared to about 50% of SI. Compared with ME (about 18%) and GT (about 20%), the signaling approach significantly improves the query efficiency. This validates our analysis in previous sections that signaling can distinguish the advantageous neighbors from the neighbors that are less so. GT is worse than SG, but is still slightly better than ME. This confirms our analysis that, due to the coordination failure in ME, the market often fails and behaves inefficiently. Obviously, according to the figure, GT overcomes the issue of coordination failures, but still suffers from the asymmetric information.

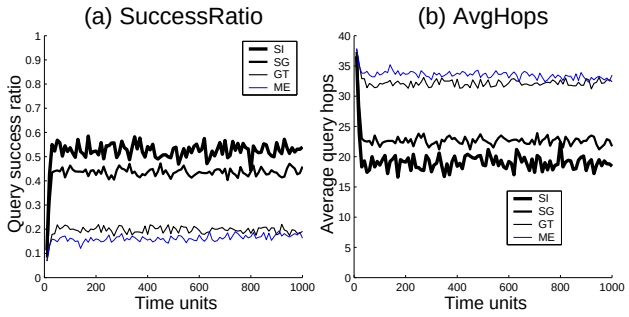

Figure 4. Query performance under uniform data popularity: (a) query success ratios; (b) average number of hops

Further, we evaluate the efficiency of queries in terms of the average number of query hops. As is evident in Fig. 4(b), the performance of SG approaches that of SI and results in much fewer query hops in average than GT and ME.

4.2.2 Simulation Results: Query Overhead

In order to evaluate the message passing overhead when performing queries, we compare all four algorithms with respect to the number of messages transmitted. Fig. 5 shows that SI causes the lowest amount of message overhead, due to the fact that the queries in SI can always find the best path to the item (*i.e.*, the shortest query path), and that there exist no transaction costs. GT and ME lead to nearly the same level of message overhead and ranks the second among the alternatives. Since in each query hop the current node needs to perform transactions with all neighbors, such an action results in additional overhead compared with that of SI. SG is evidently the worst, and causes worse levels of message passing overhead than GT and ME. This is because that, in our implementation of the signaling approach, the consumer needs to incrementally increase the price p until a producer of a high type accepts the query, which results in additional overhead. Considering the performance we have acquired by introducing signaling, the extra overhead is the cost that we have to pay, and we believe that SG has reached a better point of tradeoff between query performance and overhead.

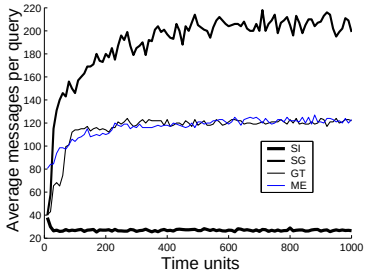

Figure 5. Query overhead in terms of the average number of messages per query under uniform data popularity

4.2.3 The case of heterogeneous item popularity

We have also evaluated all four algorithms under the assumption that the popularity of data items — the rate at which queries are issued — conforms to the *Zipf* distribution[4]. We use $\alpha = 1.2$ in our simulations, based on measurements from Saroiu *et al.* [8] on popular peer-to-peer file sharing systems.

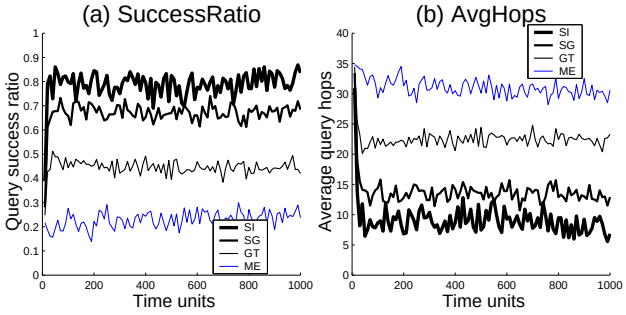

Figure 6. Query performance under Zipf-like data popularity: (a) query success ratios; (b) average number of hops

Fig. 6(a) shows that the Zipf distribution can be exploited by all four algorithms to improve query success rates, although the ranking with respect to query efficiency is still identical to that of the uniform popularity case. All four algorithms perform much more efficiently, for the reason that Zipf distribution always favors popular items and the queries on popular items encourage more nodes to create

[4]With a Zipf distribution, the popularity of the ith most popular item is proportional to $i^{-\alpha}$.

shortcuts on them. Consequently, upon one of such popular items, the number of neighbors which has the highest type $\overline{\theta}$ becomes larger and the expected type $E[\theta]$ also increases. It is encouraging to observe that, the signaling algorithm SG achieves the highest query success ratio after SI and in some cases even achieves the same success ratio as that of SI. The Zipf distribution also benefits ME and GT, whose query success ratios have nearly doubled. One interesting phenomenon is that GT performs much better in such a setting and separates itself from ME. The reason may be that, in such a setting, types of neighbors become much more diverse, multiple competitive equilibria exist in the game and GT has the advantage over ME to achieve the dominant equilibrium and to find more neighbors of high types. In addition, with respect to the average number of query hops, Fig. 6(a) shows that the same ranking is maintained as that of the uniform popularity case in Fig. 4(a).

To summarize, the ideal case of symmetric information (SI) shows that without asymmetric information, the query in a strictly unstructured peer-to-peer network should have performed much better. The signaling algorithm we have proposed (SG) encourages the nodes of high types to perform signaling actions and to reveal their private information, and therefore achieves much more efficient query performance. Another advantage of our proposed signaling algorithm is that the algorithm only incurs moderate signaling overhead compared with the ME and GT algorithms.

4.3 Discussions

We now discuss a few open but orthogonal problems with respect to our algorithms and implementations, that may possibly be addressed in future work.

First, it may require additional work to deploy our proposal to wide-area peer-to-peer network environments, most of which are beyond the scope of this paper. For example, we implicitly assume that a particular micro-payment mechanism exists as an underlying layer, which is still an open problem in the literature.

Second, it is clear that the performance of all our algorithms depends on the accuracy of type estimations, which are provided by the HU algorithm. In our simulation, we simply adopt a regular *Least Recently Used* cache replacement algorithm in the HU algorithm. We believe any improvements on the accuracy of type estimations will effectively improve the query performance of our algorithms, and they do not affect the general conclusions we have reached with respect to the nature of these algorithms.

5 Concluding remarks

Efficient peer-to-peer queries are essential to the success of unstructured peer-to-peer networks. Our observation is that, if a node can always find the best neighbor as the next hop, the query should be as efficient as flooding and at the same time causes only acceptable message overhead. Unfortunately, when selecting the next hop to forward the query, the node often suffers from the fact that the node cannot distinguish the better neighbors from the inferior candidates. This is the phenomenon of information asymmetries in selfish peer-to-peer networks, a key observation and focus of our paper. We have thoroughly examined the system behavior in asymmetric information and proposed a signaling mechanism to overcome the problem, associated with a decentralized algorithm. Our analytical and simulation results have led to and established an efficient peer-to-peer query algorithm with moderate message overhead. We believe that our studies presented in this paper form a first step towards a thorough understanding of the behavior of peer nodes in strictly selfish overlay networks, where algorithms have to be carefully designed to avoid unwarranted assumptions of cooperative behavior.

References

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

[2] G. Akerlof. The market for 'lemons': Quality uncertainty and the market mechanism. In *Quately Journal of Economics*, volume 89, pages 488–500, 1970.

[3] J. Bertrand. Theorie Mathematique de la Richesse Sociale. In *Journal des Savants*, pages 499–508, 1883.

[4] E. Cohen and S. Shenker. Replication Strategies in Unstructured Peer-to-Peer Networks. In *Proc. of ACM SIGCOMM 2002*.

[5] R. Finney. *Calculus: A complete course*. Addison-Wesley Press, 2000.

[6] 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*, 2002.

[7] A. Mas-colell, M. D. Whinston, and J. R. Green. *Microeconomic theory*. Oxford University Press, New York, 1995.

[8] S. Saroiu, P. Gnummadi, and S. Gribble. A Measurement Study of Peer-to-Peer File Sharing Systems. In *Proc. of SPIE/ACM Conference on Multimedia Computing and Networking (MMCN 2002)*.

[9] A. M. Spence. Job market signaling. In *Quately Journal of Economics*, volume 87, pages 355–374, 1973.

[10] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In *Proc. of ACM SIGCOMM 2001*.

[11] H. Zhang, A. Goel, and R. Govindan. Using the Small-World Model to Improve Freenet Performance. In *Proceedings of IEEE Infocom*, 2002.