Small-World Social Relationship Awareness in Unstructured Peer-to-Peer Networks

Hao Liao, Kuo-Chan Huang, Hung-Chang Hsiao
Department of Computer Science and Information Engineering
National Cheng-Kung University
Tainan 701, Taiwan

Abstract—Unstructured peer-to-peer (P2P) file-sharing networks are popular in the mass market. As the peers participating in unstructured networks interconnect randomly, they rely on flooding query messages to discover objects of interest. Empirical measurement studies indicate that the peers in P2P networks have similar preferences, and recently proposed unstructured P2P networks intend to organize the participating peers in a small-world (SW) fashion by exploiting the knowledge of contents stored in peers. As existing algorithms for constructing SW-based unstructured P2P networks may not precisely reveal the object sharing patterns, the resultant networks thus may not perform searches efficiently and effectively by exploiting the common interests among peers. In this paper, we suggest a novel P2P network formation algorithm to construct SW-based unstructured networks. We validate our proposal in simulations with an empirical data set, and the simulation results prove that our proposal greatly outperforms existing algorithms in terms of search efficiency and effectiveness.

Index Terms—Peer-to-peer systems, unstructured overlay networks, social relations

I. INTRODUCTION

Peer-to-Peer (P2P) networks (or overlay networks) have been widely deployed in the Internet, and they provide various services such as file sharing, information retrieval, media streaming, and telephony. P2P applications are popular as they primarily provide low entry barriers and self-scaling. Recent studies (e.g., [1], [2]) reveal that P2P applications dominate file sharing and information retrieval. Gnutella [3] is a popular P2P search protocol in the mass market. Specifically, as Gnutella networks are unstructured, and the peers participating in networks connect to one another randomly, the peers search objects in the networks by message flooding. To flood a message, an inquiry peer broadcasts the message to its neighbors (by the neighbors of peer i, we mean those peers that have end-to-end connections with i). The broadcast message is associated with a positive integer time-to-live (TTL) value. Upon receiving a message, the peer (say, j) decreases the TTL value associated with the message by one and then relays the message with the updated TTL value to its neighbors, except the one sending the message to j, if the TTL value remains positive. In addition to forwarding the message to the neighbors, j searches its local store to see if it can provide the objects requested by peer i. Conceptually, if j has the requested objects and is willing to supply the objects, then j either directly replies i the objects, or returns the objects along the overlay path at which the query message traverses from i to j.

In this study, we are interested in optimizing the search performance in Gnutella-like unstructured P2P networks. Existing orthogonal techniques in the literature for improving search performance in unstructured P2P networks include indexing, replications, superpeer architectures, and overlay topologies, among others. We primarily study in this paper the overlay topology formation technique for unstructured P2P networks. In particular, as recent measurement studies (e.g., [4]) show that peers exhibiting similar preferences in a P2P network are likely to resolve the queries issued by their peers in the community, we aim to organize the participating peers to exploit their social community structures implicitly inherited from the objects shared in the network. Hereafter, the peers having similar interests are informally named similar peers for brevity.

The study [5] of Zhu et al. presents a distributed algorithm for constructing unstructured P2P networks with the small-world (SW) social feature. Consider an unstructured P2P network represented by $G = (V, E)$, where $V$ represents the set of peers (or nodes) in the network, and $E$ denotes the set of existing links that connect the nodes in $V$. $G$ is an SW-based network if it possesses a low diameter and high cluster coefficient [6]. The diameter measures the maximum shortest path length (in terms of overlay hopcount) between any two nodes in $V$, while the cluster coefficient of any node $v \in V$ presents the density of existing connections between the neighboring nodes of $v$. Typical random graphs have a small diameter but low cluster coefficient [7]. In contrast, an SW-based network allows rapid and reliable message delivery as any two nodes in the network have a low path length due to their low diameter. Further, the neighbors of any node are likely to connect to one another due to the high cluster coefficient. Zhu et al. in [5] suggest that similar peers are connected to one another, while each peer additionally maintains a few random neighbors, thus resulting in an SW-based P2P network. Later, studies such as [8] also present algorithms for forming SW-based unstructured P2P networks.

In this paper, we present a novel network formation algorithm to construct SW-based unstructured overlays and validate our proposal in simulations using empirical data...
set. In addition, our proposal is compared to state-of-the-art solutions, namely, GES [5] and SocioNet [8]. The simulation results conclude that our SW-based network clearly outperforms prior solutions in terms of search efficiency and effectiveness. In particular, our study reveals that existing SW-based unstructured P2P networks (e.g., [5], [8]) implicitly mix random graphs in which an overlay link connecting any two peers is in a probability of \( O(n^{-1}) \) [7], resulting in inefficiency and ineffective search performance. In contrast, our proposed unstructured overlays take advantage of small-world feature exhibited by the objects sharing pattern, enabling to progressively and effectively exploit the similarity of participating peers for searches towards the destinations.

The remainder of our paper is organized as follows. Section II discusses related works. Section III introduces our proposed SW-based networks based on our analytical model. Section IV assesses the performance of our solution and compares our proposal to prior algorithms. Finally, Section V summarizes our study with possible future research directions.

II. RELATED WORK

Kleinberg presents an algorithm that constructs SW-based P2P networks for unicast routing [9]. In [9], the peers in the network are organized in a \( k \)-dimensional lattice, where each peer in the network has a unique coordinate. To route a message to a destination peer, the source peer needs to specify the coordinate of the target location. Upon receiving a message, the intermediate peer relays the message to one of its neighbors in the lattice, which is closest to the destination. Kleinberg assumes in [9] that the participating peers in the network are static without freely joining and departing. In addition, there is a central entity in the network to compute the neighbors for each of the participating peers. Unlike [9], distributed algorithms that create SW-based P2P networks for unicast-based routing can be found in the literature (e.g., [10]), allowing peers to arbitrarily join/leave the system and construct their neighbors independently without relying on a central entity. While [9], [10] present SW-based P2P networks for unicast-based routing, our proposal in this paper aims at designing SW-based P2P networks in which the participating peers depend on flooding-based P2P search protocols in the mass market (e.g., Gnutella [3]). Specifically, in contrast to [9], [10] that offer “semantic-free” key-based search, our SW-based P2P network supports complex search (e.g., query by keywords) by nature. We present in this paper how our proposed P2P network exploits the content knowledge of the participating peers for an efficient and effective search.

pSearch [11] and SSW [12] are content-based P2P networks providing semantic search. Similar to most P2P networks based on distributed hash tables (e.g., Chord [13]), in pSearch and SSW, each published object, represented by a latent semantic vector [14], needs to be first indexed into the network in which the participating peers are formatted in a well-structured manner and host a disjoint key subspace. Therefore, the participating peers need to maintain foreign indices, the indices of objects stored in remote peers. To locate an object, a requesting peer routes a message towards the peer responsible for the key subspace wherein the object is indexed. On the contrary, in our study, we present a construction of unstructured P2P networks where the participating peers need not organize themselves into a rigid, deterministic topology structure, avoiding the maintenance overhead of overlay topology. Unlike pSearch [11] and SSW [12], the peers in our proposed network host the objects of interest and maintain no foreign indices, eliminating storage and bandwidth overheads for publishing and managing such indices.

Perhaps, the studies most relevant to ours are GES [5] and SocioNet [8]. Both GES and SocioNet maintain no foreign indices for objects published in the system, and they essentially rely on message flooding to discover requested objects. As peers having similar preferences are likely to provide objects to one another [4], in GES by Zhu et al. [5], similar peers are clustered. In addition, each peer in GES connects to a number of peers selected uniformly at random from the system. Unlike GES, SocioNet [8] requires benchmarking an operational overlay network globally in order to determine the rewiring probability for each overlay connection. On one hand, our proposal presented in this paper operates pragmatically in the sense that each participating peer depends on its local knowledge to select its overlay neighbors, thus adapting a large-scaled, dynamic environment. On the other, we also show through extensive simulations that there is no short path in GES and SocioNet that exploits similar peers in order to discover the requested objects.

III. OUR PROPOSAL

A. Object-Set similarity Map

Consider a set of objects, \( O \). We first define the following:

**Definition 1.** The object-set similarity function \( F: 2^O \times 2^O \rightarrow \mathbb{R}^+_0 \)

measures the degree of similarity between any two sets of objects \( O_u \subseteq 2^O \) and \( O_v \subseteq 2^O \).

To simplify the discussion, we let \( F(O_u, O_v) = F(O_v, O_u) \), which is symmetric, in this paper. For example, in a P2P network, \( F(O_u, O_v) = \frac{|O_u \cap O_v|}{|O_u \cup O_v|} \) describes the percentage of common objects appearing in the both object sets, \( O_u \) and \( O_v \), in peers \( u \) and \( v \), respectively. \(^1\)

**Definition 2.** An object-set similarity map \( \mathcal{G} = (2^O, \mathcal{E}) \) is a graph where \( 2^O \) denotes the universal set of all possible subsets of objects (or vertices) in \( 2^O \), and \( \mathcal{E} \) is the set of edges, where each edge \( (O_u, O_v) \in \mathcal{E} \) indicates that vertices \( O_u \subseteq 2^O \) and \( O_v \subseteq 2^O \) are similar to some extent.

In \( \mathcal{G} \), \( \mathcal{E} \) depends on the object-set similarity function, \( F(\cdot) \). More specifically, as we assume that \( F(\cdot) \) is symmetric, an edge \( (O_u, O_v) \) appears in \( \mathcal{E} \) if and only if \( F(O_u, O_v) = F(O_v, O_u) > 0 \). Notably, (1) \( F(\cdot) \) is symmetric, and \( \mathcal{G} \) is

\(^1\)Possibly, \( F(O_u, O_v) \) is defined as the inverse of the cosine angle of two “summarized” latent semantic vectors representing any two peers \( u \) and \( v \) in a P2P network, where each element in a summarized vector for any peer \( i \) calculates the total frequency of the corresponding keyword appearing in the data items stored in \( i \) [5].
thus an undirected graph. (2) $G$ may not be a connected graph.

(3) We explicitly differentiate the vertices in an object-set similarity map and the peers (nodes) in a P2P network. We refer the vertices to the peers when we discuss the optimization of searching in a P2P network that exploits its object-set similarity map.

Any subgraph of an object-set similarity map $G = (2^O, E)$ is denoted by $\mathcal{G} = (\mathcal{P}, \mathcal{L})$ if and only if $\mathcal{P} \subseteq 2^O$ and $\mathcal{L} \subseteq E$. (By definition, $G$ itself is also a subgraph.) For instance, the object sharing pattern in the unstructured P2P network in the mass market, that is, Gnutella, is essentially a subgraph of $G = (2^O, E)$, where $O$ is the set of all files shared by the peers in the Gnutella network. We do not explicitly differentiate in this paper the peers maintaining an identical set of objects as this is unlikely to occur in a real setting.

![Fig. 1. An example of a subgraph $G = (P, L)$ in the object-set similarity map $G = (2^O, E)$, where $O = \{a, b, c\}$ and thus $O_1 = \{a\}$, $O_2 = \{a, c\}$, $O_3 = \{c\}$, $O_4 = \{b, c\}$, $O_6 = \{a, b\}$, and $O_7 = \{a, b, c\}$. Here, $\mathcal{F}(O_4, O_5) = \left|O_4 - O_5\right|$, $\forall u \in \mathcal{P} = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7\}$.](image)

Consider any subgraph $G = (P, L)$, we further define the following metric:

**Definition 3.** Given $G = (P, L)$, the object similarity distance between two distinct vertices $O_u \in \mathcal{P}$ and $O_v \in \mathcal{P}$, denoted by $D(O_u, O_v)$, is defined as the minimum number of edges in $L$ that connect $O_u$ and $O_v$. Without loss of generality, we let $D(O_u, O_u) = 0$ if $u = v$.

For example, in Fig. 1, $D(O_1, O_3) = 2$ and $D(O_3, O_7) = 1$.

We remark that since any subgraph $G = (P, L)$ of $G$ is an undirected graph, $D(O_u, O_v) = D(O_v, O_u)$. Second, $D(O_u, O_v) \geq 1$. Third, for any three vertices $O_u$, $O_v$, and $O_w$ in $P$, if $D(O_u, O_v) < D(O_u, O_w)$, then we say that $O_v$ is more similar to $O_u$ than $O_w$.

**Definition 4.** The scope of a vertex $O_u \in \mathcal{P}$ in $G = (P, L)$ within a given similarity distance $d$ is defined as

$$S_{O_u}(d) = \{O_v | D(O_u, O_v) \leq d, \forall O_v \in \mathcal{P}\}. \quad (2)$$

Therefore, the cardinality of the scope $S_{O_u}(d)$, $|S_{O_u}(d)|$, measures the number of vertices in $S_{O_u}(d)$. Consequently, the vertices in $S_{O_u}(d_1)$ are similar to one another compared with those in $S_{O_u}(d_2)$ if $d_1 < d_2$.

In Fig. 1, $S_{O_1}(1) = \{O_1, O_2, O_6, O_7\}$ and $S_{O_1}(2) = \{O_1, O_2, O_3, O_4, O_6, O_7\}$, for example. Thus, $|S_{O_1}(1)| = 4$ and $|S_{O_1}(2)| = 6$.

**B. Problem Formulation**

Consider any given unstructured P2P network $G = (V, E)$, where $V$ is the set of participating peers, and $E$ is the set of overlay connections linking the peers in $V$. The peers in $G$ are interconnected randomly. As the object sharing pattern in $G$ is essentially an object-similarity map, $\mathcal{G}$, our goal is to restructure $G$ in order to satisfy the following objectives:

1) $G$ can match $\mathcal{G}$

2) $G$ shall be augmented so that a flooded query message can rapidly visit the peers having objects similar to the request.

In the following discussions, we denote $I_v$ as the set of nodes that peer $v \in V$ connects with in order to match $\mathcal{G}$, while $\Phi_v$ includes $v$’s extra connections to minimize the hopcount of routing each of the flooded messages. We let $B_v = I_v \cup \Phi_v$ be the target set of peers that $v$ in our proposal shall link to.

**C. Matching $G$ with $\mathcal{G}$**

Obviously, any peer $v$ in $G = (V, E)$ shall link to the nodes in $S_v(1)$ so that the reshaped $G$ can match its object-set similarity map $\mathcal{G}$. Clearly, in $G = (V, E)$, the set of neighboring nodes of $v$, that is, $I_v = \{u | (v, u) \in E\}$, is unlikely to be identical to $S_v(1)$. This means $\sum_{u \in I_v} F(v, u) < \sum_{u \in S_v(1)} F(v, u)$. Consequently, we reformat $G$ such that for any $v \in V$, $\sum_{u \in I_v} F(v, u)$ is maximized. Therefore, we solve the optimization problem as follows:

$$\text{max} \sum_{v \in V} \sum_{u \in V - \{v\}} x_{vu} F(v, u) \quad (3)$$

s.t. $$|I_v| \leq \text{MAX}_{I_v} \quad (4)$$

$$\sum_{(v, u) \in \Phi(v)} x_{vu} \geq 1 \quad \forall \Phi \neq \emptyset \subset V \quad (5)$$

$$x_{vu} \in \{0, 1\} \quad \forall \Phi \neq \emptyset \subset V \quad (6)$$

where $x_{vu}$ is a binary variable indicating whether the overlay link $(v, u)$ shall appear in the reshaped $G$ or not. $|I_v|$ in Eq. (4) is the number of connections that peer $v$ maintains for its $S_v(1)$. Here, $I_v$ shall be no more than $\text{MAX}_{I_v}$, the maximum number of connections $v$ can maintain in its $I_v$. In a practical P2P system, each peer can only maintain a limited number of connections, considering its capacity. $\Phi(v)$ consists of all edges connecting two components $V$ and $V - \{v\}$.

Eq. (4) indicates that for any subset $V \subset V$, there is at least one overlay link connecting the two components $V$ and $V - \{v\}$, ensuring the connectivity of the reshaped $G$.

To solve the above problem, we rely on the Monte Carlo method [15]. Let $G' = (V', E')$ be the resultant reshaped $G$, $T_G = \sum_{(v, u) \in E} F(v, u)$ and $T_{G'} = \sum_{(v, u) \in E'} F(v, u)$. Our proposal works as follows. If $T_{G'} < T_G$, $G'$ is accepted; otherwise, we accept $G'$ with a probability of $\frac{e^{\frac{T_{G'}-T_G}{\alpha}}}{e^{\frac{T_{G'}-T_G}{\alpha}}}$. Here, $T$
is a system parameter. More specifically, in our proposal, each node \( v \in V \) independently performs the algorithm as described previously, where \( I_v = \{ w | u \in \mathcal{I}_u, u \in I_v \} - \mathcal{I}_v - \{ v \} \).

Note that (1) in our proposal, \( v \) collects and maintains its local knowledge (i.e., \( \mathcal{I}_v \) and \( I_v \)) to perform the algorithm. (2) Node \( v \) performs the above mentioned algorithm iteratively. Given \( G \), \( v \) improves \( G \), resulting in \( G' \). Then let \( G = G' \), and \( v \) repeats the algorithm for such \( G \). According to [15], if each node performs our algorithm for a sufficient time, then \( G' \) is constructed with the probability proportional to \( e^{-\frac{\alpha}{|S|}} \).

As a result, our proposal constructs \( G' \) to have a large \( T_{G'} \) in a higher probability compared with those \( G' \)'s with a relatively smaller \( T_{G'} \). (3) If peer \( v \) connects to peer \( w \in I_v \), then it will remove an existing connection with \( u \in \mathcal{I}_v \). This is because our implementation does not intend to increase the total number of connections in \( G \) so that \( G \) will not be overwhelmed in maintaining the sheer number of overlay links due to our optimization process.

### D. Building Extra Connections

In our proposal, each peer \( v \in V \) shall create a number of extra overlay links by connecting to a set of nodes, \( \Phi_v \), each \( t \in \Phi_v \), selected in a probability of \( \Pr(v; t) \). \( v \) first issues a random walker to sample \( t \) uniformly at random from network \( G \). Once \( t \) is sampled, \( u \) invites \( t \) as its neighbor with a probability of \( \Pr(v; t) \). \( \Pr(v; t) \) is given in Section IV.

Accordingly, we discuss how \( t \) can be sampled uniformly at random from the network. Consider any peer \( x \) in the network \( G \). Upon receiving the walker originated by peer \( v \), \( x \) dispatches the walker to one (say, peer \( y \)) of its overlay neighbors according to the following probability:

\[
T(x, y) = \begin{cases} \varphi_x \min \left( \frac{1}{\deg_y - 1}, \frac{1}{\deg_x} \right) & \text{if } y \in B_x, \\ 0 & \text{if } y \notin B_x, \end{cases}
\]

where \( \varphi_x \) is an arbitrarily positive value such that \( T(x, x) = 1 - \sum_{y \in B_x} T(x, y) > 0 \), and \( \deg_y \) denotes the number of connections maintained by peer \( y \) in \( G \). Here, \( \deg_x \) includes the number of extra connections in addition to \( |\mathcal{I}_x| \).

Notably, Eq. (7) allows a unique stationary probability distribution such that \( \nu v \) selects any peer \( t \) from \( V \) with a probability of \( \pi(t) = \frac{1}{N} \), where \( N \) is the total number of active peers in the system. This is because (i) the Markov chain defined by the probability transition matrix \( [T(x, y)] \) is aperiodic (each node has a positive probability to pick itself), (ii) irreducible (there is a positive probability to pick any node from \( V \)), and (iii) time reversible (\( \pi(x)T(x, y) = \pi(y)T(y, x) \) for all \( x \neq y \in V \) [16].

In summary, given \( G \), \( G \) is matched with its object-set similarity map \( \mathcal{G} \) due to our algorithm mentioned in Section III-C, resulting in the unstructured network as \( G' = (V, E') \). While \( G' \) is improved over time by having each peer perform the algorithm in Section III-C iteratively, we augment \( G' \) with extra connections, namely, \( E_{\text{extra}} = \cup_{v \in V} \{ (v, u) | u \in \Phi_v \} \) as discussed in this section, leading to our target overlay \( G^* = (V, E' \cup E_{\text{extra}}) \). Notice that each peer \( v \) in \( G^* \) manages its overlay connections for \( E' \) and \( E_{\text{extra}} \) independently.

Second, each random walk originated by any peer \( v \) samples peers from \( G^* \) for creating its \( \Phi_v \).

### IV. Simulations

#### A. Experimental Setup

We have developed an event-driven simulator to evaluate various SW-based unstructured P2P networks. In the simulations, the P2P network to be optimized is modeled as a random regular graph [7], denoted by \( G = (V, E) \), where each peer in \( V \) has the default number of overlay connections equal to 4. Given a \( G \), we investigate the performance of the state-of-the-art unstructured SW-based search overlays, namely, GES [5] and SocioNet [8], and compare these overlays to our proposal.

In the simulations, each peer \( v \) can maintain \( B_v \leq 16 \) overlay links, including incoming and outgoing ones. Since each peer \( v \) in GES, SocioNet, and our proposal attempts to link to the most similar peers, \( v \) can connect with such neighbors up to \( \text{MAX}_{\mathcal{I}_v} = 4 \). Additionally, each peer \( v \) in GES and SocioNet links to \( \Phi_v \), extra neighbors, where \( \Phi_v = 5, 10 \).

In our study, the default value of \( \Phi_v \) is 5 as the number of peers simulated is small. Notably, in our proposal, we let \( T = 0.01 \) (see Section III-C).

Given a random regular graph, we perform \( R = 200 \sim 51,200 \) algorithm rounds for each of GES, SocioNet, and our proposal. We measure the performance metrics (discussed later) after performing \( R \) rounds of each investigated algorithm. The performance results we report in the following sections are averaged over 50 extensive simulation runs. More specifically, we generate 50 random regular graphs as the input \( G \)'s of GES, SocioNet, and our proposed P2P network. For each of the 50 \( G \)'s, we perform GES, SocioNet, and our proposal; then we average the performance metrics measured after \( R \) algorithm rounds.

We model the social relations among the peers with an empirical data set, i.e., \( \text{YOUTUBE} \), and represent the social relations with the object-set similarity map discussed in Section III-A. In \( \text{YOUTUBE} \), each shared video file, \( f \), is linked to a number of distinct videos. Specifically, \( f \) connects to at most 20 distinct videos in the trace [17], resulting in the object-set similarity map \( \mathcal{G}_f \) for \( \text{YOUTUBE} \). We let \( \mathcal{O} \) be the set of published video files, each individual file \( \{ f \} \) be a vertex, and \( \mathcal{F}(\mathcal{O}_w, \mathcal{O}_v) \) be 1 if a video \( w \) links to another video \( v \) in the trace; \( \mathcal{F}(\mathcal{O}_w, \mathcal{O}_v) = 0 \), otherwise.

The distribution of \( |S_{\text{y}}(k)| \) for \( \text{YOUTUBE} \) is illustrated in Fig. 2. We also depict \( |S_{\text{y}}(k)| \) for a subgraph \( \mathcal{G}_Y \subset \mathcal{G}_f \) in
the plot, where each vertex in $G_Y$ connects to 8 vertices at most. As shown in Fig. 2, the distribution for $|S_o(k)|$ in $G_Y$ resembles Pareto distribution $\beta x^\alpha$ with $\alpha \approx 3.26$ and $\beta \approx 0.35$.

In our simulations, for the object-set similarity map, we assign each simulated peer a unique vertex selected randomly without replacement from the map. Due to the scalability of our simulator, we study up to around 27,000 peers for YouTube.

Notably, as mentioned in Section III-D, each peer $v$ links to at most $\Phi_v$ extra neighbors each picked with a probability of $P_{r(v,t)}$. In this study, we let $P_{r(v,t)}$ be proportional to $1/|D(v,t)|$. Due to space limitation, we show, through a rigorously mathematical analysis, in our accompanying paper [18] that if the scope of each peer $v$, i.e., $|S_v(k)|$, exhibits the power-law distribution as plotted in Fig. 2, then $P_{r(v,t)}$ shall be proportional to $1/|D(v,t)|$ such that query messages can efficiently, effectively and progressively exploit the similarity of peers towards those peers capable of resolving the queries. Additionally, we present in [18] how each peer $v$ in our proposal measures $D(v,t)$ and thus computes $P_{r(v,t)}$.

B. The Simulation Results

1) Precision: Consider an object $o \in \mathcal{O}$ in the system. Let $C_o$ be the set of peers that replicates the object, $o$. Consider a peer, $v$, that issues a query, denoted by $Q(v,o)$, in searching for $o$. Let $P_q$ be the set of peers that receive the query message $Q(v,o)$. Among the peers in $P_q$, denote the set $P_o \subseteq P_q$ as the peers that store the object $o$, that is, $P_o = C_o \cap P_q$. Then the performance metric precision is defined as $\frac{|P_o|}{P_q}$. Clearly, the precision metric reveals the ratio of discovered objects to the search cost.

We define the metric, $r$, as the ratio of the number of peers replicating a given object $o \in \mathcal{O}$ to the total number of peers, $N$, in the system. Consider an object $o$ published by peer $v$. Given a replication ratio $r$, we replicate $o$ to the set of peers $P_v$, with each $u \in P_v$ having a similarity distance to $v$ no more than $d$ such that $r = \frac{|\{u \in P_v : \alpha(u,v) < d\}|}{N}$.

In the experiments discussed in this section, we study two types of queries. One type we consider is letting each peer query an object $o \in \mathcal{O}$ with a probability of $\frac{1}{|\mathcal{O}|}$. For the second type, each peer $v$ searches an object $o$ selected from the universal object set $\mathcal{O}$ with a probability proportional to $\frac{1}{|\mathcal{O}|} \times |S_o(k)|$, where peer $u$ caches the searched object. In contrast to the former type, the latter type models the social relationships among the peers, that is, similar peers are likely to search objects stored in one another. Note that we let $\Phi_v = 10$ in the experiments discussed in this section such that the GES network is connected.

Figs. 3 and 4 plot the measured precision against the processing cost for the first and second types of queries, respectively. We investigate $r = 0.39\%, 3.3\%$ and $36\%$ in this experiment. By the processing cost, we mean the ratio of the number of peers receiving an identical query to $N$. The simulation results conclude that our proposal constantly outperforms GES and SocioNet in terms of precision. In particular, for the first type of query, our proposal outperforms GES by 10\% and SocioNet by 8\% if $r = 3.3\%$ and the processing cost is 20\%. However, ours is better than both GES and SocioNet by 15\% in case $r = 3.3\%$ and the processing cost is 20\% for the second type of query (Fig. 4(b)). In contrast to GES and SocioNet mix random graphs in the networks (where each peer links to another peer with a probability of $\frac{1}{N}$), our proposal is more effective than GES and SocioNet in exploiting the social relations among peers.

2) Overlay Similarity Distance: For an insight into the quality of overlay topologies of GES, SocioNet and our proposal, we measure the averaged similarity distance for the $k^{th}$-hop neighbors with respect to any node in the network. By the $k^{th}$-hop neighbors of any node $v$, we mean those peers, denoted by the set $H_v(k)$, that have the shortest overlay paths of length $k$ to $v$. Therefore, the averaged similarity distance for the $k^{th}$-hop neighbors with respect to node $v$ is $D_v(k) = \frac{\sum_{u \in H_v(k)} D(v,u)}{|H_v(k)|}$. The performance metric, $D_v(k)$, helps understand how a query message exploits similar peers in $H_v(k)$ before it reaches the destination node, $v$. If there exist peers $u_k \in H_v(k), u_{k-1} \in H_v(k-1), \cdots, u_1 \in H_v(1)$ on a path $u_k \rightarrow u_{k-1} \rightarrow \cdots \rightarrow u_1$ to $v$ and $D(v,u_k) > D(v,u_{k-1}) > \cdots > D(v,u_1)$, then the query message discovers $v$ by effectively taking the social relations among the peers in $\bigcup_{i=1}^{k} H_v(i)$.

Fig. 5 illustrates the simulation results for $D_v(k)$. Here, we average $D_v(k)$ for all nodes $v$ in the system, given a $k$ value (where $k = 1, 2, 3, \cdots$). The simulation results reveal that our proposal performs very well, enabling a query message to effectively exploit the peers similar to the requested node, that is, $D_v(i) > D_v(j)$ and $i > j$. On the contrary, a query message in GES and SocioNet cannot effectively take advantage of the social relations among the peers in order to approach the nodes having the requested objects.

V. SUMMARY AND FUTURE WORK

We have presented a novel SW-based unstructured P2P network, intending to take advantage of the social relations among participating peers for enhancing search performance. Through extensive simulations with an empirical data set, we have validated our proposed network. In addition, the simulation results conclude that our proposal outperforms GES and SocioNet by effectively exploits the social relations among peers. As our emphasis in this study is to develop an overlay interconnect protocol for enhancing searches by exploiting the social relations among peers, we will investigate optimization techniques such as the object indexing, placement, and replication, which are orthogonal to our proposed overlay infrastructure in the future.

REFERENCES

Fig. 3. The precision ((a), (b), and (c)) for GES, SocioNet, and our proposal with $G_Y$ by varying the replication ratio $r$ (queries to objects picked uniformly at random).

Fig. 4. The precision ((a), (b), and (c)) for GES, SocioNet, and our proposal with $G_Y$ by varying the replication ratio $r$ (queries to objects picked non-uniformly at random).

Fig. 5. The averaged similarity distance for $k^{th}$-hop neighbors with respect to any node for $G_Y$.


