Expansion and Search in Networks

Arun S. Maiya
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, Illinois 60607
amaiya@cs.uic.edu

Tanya Y. Berger-Wolf
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, Illinois 60607
tanyabw@cs.uic.edu

ABSTRACT

Borrowing from concepts in expander graphs, we study the expansion properties of real-world, complex networks (e.g. social networks, unstructured peer-to-peer or P2P networks) and the extent to which these properties can be exploited to understand and address the problem of decentralized search. We first produce samples that concisely capture the overall expansion properties of an entire network, which we collectively refer to as the *expansion signature*. Using these signatures, we find a correspondence between the magnitude of maximum expansion and the extent to which a network can be efficiently searched. We further find evidence that standard graph-theoretic measures, such as average path length, fail to fully explain the level of "searchability" or ease of information diffusion and dissemination in a network. Finally, we demonstrate that this high expansion can be leveraged to facilitate decentralized search in networks and show that an expansion-based search strategy outperforms typical search methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Algorithms; Experimentation, Measurement

Keywords

expansion, decentralized search, P2P, peer-to-peer networks, social network analysis, complex networks, graph mining, expander graphs, focused web crawling, MANET

1. INTRODUCTION

Motivated by the algorithmic problem of searching large graphs, we study the expansion properties of real-world networks and the extent to which these properties can be exploited to better understand and facilitate decentralized search.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

Complex networks of linked entities arise across diverse domains, from sociology (e.g. social networks) to biology (e.g. neural networks, protein interactions) to technological and information systems (e.g. P2P networks, the Web, power grids). Scientists from disparate fields, especially within the past decade, have attempted to characterize both the structure and function of these networked systems. The standard approach here has been to measure topological features of the graph representing the network and correlate them with functional or dynamic aspects of the network (e.g. evolution of the network, the behavior of processes that occur over the network). For instance, in their seminal paper on small-world networks, Watts and Strogatz [33] showed that many real-world networks simultaneously exhibit short average path lengths and relatively high degrees of clustering. They further showed that these features can facilitate spreading processes across a network (e.g. the spread of a virus) [33]. In this work, we study a feature that has received comparatively less attention in the study of real-world[1] networks: *expansion*.

1.1 Expansion

Given a network $G = (V, E)$ where V is the set of nodes and E is the set of links among the nodes, the *expansion* of a set of nodes $S \subset V$ is a function of the number of nodes in $V - S$ to which S is connected (see Section 3 for more precise definitions). That is, if $N(S)$ is the set of nodes to which S is connected, then the *expansion* of S is $\frac{|N(S)|}{|S|}$. Informally, an expander graph is a graph in which any subset of nodes has good expansion (i.e. has many neighbors) [12]. For instance, a network is said to be a γ-expander if every $S \subset V$ has an expansion of *at least* γ, where $|S| < \frac{|V|}{2}$ [12]. Thus, the classical definition of an expander graph focuses on samples with the *minimum* expansion (as γ is a minimum if every sample S has an expansion of at least γ). Expander graphs have been shown to have many applications from constructing error correcting codes to routing calls in telephone switching networks [12]. Although these classical expander graphs are well-studied, there has been surprisingly less attention paid to studying the sample in a network with the *maximum* expansion, and it is this in which we are most interested. Our overall aim in this work is to investigate

[1]There is a large body of work studying expansion in theoretical computer science and graph theory. However, much of this work focuses on 1) synthetic graphs that do not normally arise in the real-world such as d-regular graphs and 2) the *minimum* (not maximum) expansion in these graphs [12].

the extent to which a vertex set with high expansion can be leveraged to both understand and facilitate efficient decentralized search in networks. We now describe the problem of decentralized search.

1.2 Search

A central algorithmic problem in the study of complex networks is how to efficiently search them in a decentralized manner. This scenario arises in many practical applications from querying peer-to-peer file sharing networks to focused Web crawling [17]. Starting from some initial source node, we must locate, access, or route a message to some other target node in the network. Without full knowledge of the global network topology, we are unable to simply compute the shortest path or access the target node directly. Thus, we must hop from node to node until the target node is found. Decentralized search, then, is related to information diffusion or dissemination, as an efficient search will involve efficiently disseminating a query message to large portions of the network. In this work, we study the effect of network expansion on decentralized search. If, using only local information, nodes are visited in such a way that their overall distance is close to many other nodes (i.e. the set of visited nodes has high expansion), then the efficiency of search might be improved. Moreover, the magnitude of expansion in a network may shed light on the extent to which a network can be efficiently searched by *any* search algorithm. These are precisely the questions we investigate here.

1.3 Contributions and Summary of Findings

We rigorously investigate decentralized search across a span of networks much wider and more diverse than previously studied in work on searching graphs. Our main contributions include the following:

- Borrowing from concepts in expander graphs, we introduce the concept of *expansion signatures*, which concisely captures the overall expansion properties of a network. We find that, in many networks, relatively small samples of nodes can exhibit significantly high expansion. However, we also find that there are a few networks for which small samples with high expansion may not exist.

- We propose an expansion-based, decentralized search strategy, which explicitly tries to locate these samples with high expansion in order to quickly discover the most nodes in a search. We evaluate a number of search algorithms such as degree-based searches, breadth-first searches, and random walks. We show that an expansion-based search strategy generally outperforms others.

- We demonstrate that *expansion signatures* correctly infer the extent to which a network can be efficiently searched (which we refer to as "searchability"). Moreover, we show that it is the maximum expansion in a network, rather than the minimum expansion, that contributes most to searchability and information dissemination in a network. At the same time, we find that standard graph-theoretic measures, such as average path length, *fail* to fully explain the extent to which a network is easily searched.

The last point is our most significant finding. Existing works have mostly studied the effect of *minimum* expansion on the ease of dissemination in a network (e.g. [5,12]). Other works have focused on the effect of standard graph properties such as average path length on searchability and ease of dissemination (e.g. [13]). For the first time, we show that it is the *maximum* expansion (rather than minimum expansion) that most affects efficient searchability. Our results, then, offer a more comprehensive picture of decentralized search and information diffusion in networks than has previously been appreciated.

2. BACKGROUND AND RELATED WORK

Interestingly, one of the first experiments on decentralized search in networks was the famous chain-letter study by the social psychologist Stanley Milgram [25]. In this experiment, participants were given the name, address, and occupation of an unknown target person and told to forward a chain letter to this person by passing the letter on to a single acquaintance meeting two main conditions: 1) the acquaintance must be someone with whom the individual knew on a first-name basis and 2) the acquaintance chosen should be the one perceived as closest to the target [25]. This study not only provided some evidence for short paths in social networks[2] (the median path length between sources and targets was 6 among letters that reached their destination), but also showed that individuals were able to collectively discover these short paths *without* full knowledge of the network [16]. Kleinberg [16] later modeled this problem algorithmically using a 2-dimensional grid with probabilistically-added long-range connections and shed light on the precise conditions that these short paths were discoverable using a decentralized search algorithm.

In both the works of Milgram [25] and Kleinberg [16], although nodes had no knowledge of global network connectivity, there was, in fact, *external* knowledge that aided searches. In the Milgram experiment, for instance, individuals were instructed to forward letters to the local acquaintance that was perceived as being closest to the target - as measured by geographic or occupational similarity in many cases. Thus, external knowledge of geographical distance and occupational similarity was employed as an aid in the search heuristic. In other words, there was knowledge (or, at least, an assumption) that individuals closer geographically or more similar occupationally are more likely to know each other. Liben-Nowell et al. [23], in fact, showed some evidence for the geographical basis of online friendships in the LiveJournal social network. In Kleinberg's model also, it is assumed that message holders have access to external knowledge and know the local contacts of *all* nodes in the network (i.e. nodes are aware of the Manhattan distances between all nodes in the underlying grid structure and use this information as a forwarding heuristic). Boguna et al. [4] have recently modeled this external information as a hidden metric space.

Unfortunately, in many real-world scenarios, such as unstructured peer-to-peer file sharing networks, these types of

[2]As noted in [18], a number of issues exist in Milgram's results. For instance, many chain letters failed to ever reach the target. Nevertheless, the conclusion that short path lengths exist in social networks is generally accepted today and has been verified in many networked data [17].

external information and similarity-based heuristics are unavailable as search aids, and the problem of decentralized search becomes even more challenging. Typical approaches here resort to variations on flooding the network (which can be unscalable), random walks (which may be less effective in finding information), or imposing structure on the network to improve searchability (which requires additional overhead) [31]. For a review of decentralized search both in the contexts of complex networks and specifically P2P, one may refer to [17, 26, 31]. Our focus in this work is to investigate efficient search on networks with *arbitrary* structure in which similarity-based heuristics (e.g. geographic distance) are unavailable. For the first time, we investigate the relationship between *expansion* and the extent to which a network is efficiently searchable. We further show that a search strategy based on expansion generally outperforms typical existing approaches such as random walks and flooding-based techniques. We begin a discussion of our work with some preliminaries.

3. PRELIMINARIES

3.1 Notations and Definitions

We now briefly describe some notations and definitions used throughout this paper.

Definition 1. $G = (V, E)$ is an undirected *network* or *graph* where V is a set of vertices (or nodes) and $E \subseteq V \times V$ is a set of edges (or links between the nodes). We will use the terms *network* and *graph* interchangeably.

Definition 2. A *sample* S is a subset of vertices, $S \subset V$.

Definition 3. $N(S)$ is the *neighborhood* of S if $N(S) = \{w \in V - S : \exists v \in S \; s.t. \; (v, w) \in E\}$. The *neighborhood* may also be referred to as the *frontier* of a sample S.

Definition 4. The *expansion* of a sample[3] S is:

$$\frac{|N(S)|}{|S|}$$

Definition 5. The *maximum expander set* of size k is a sample S of size k with the maximal expansion:

$$\underset{S:|S|=k}{\operatorname{argmax}} \frac{|N(S)|}{|S|}$$

Definition 6. The *expansion quality* of a sample S is the normalized[4] *expansion*:

$$\frac{|N(S)|}{|S|} \div \frac{|V-S|}{|S|} = \frac{|N(S)|}{|V-S|}$$

.

Notice that, given a sample S, the maximum possible expansion on *any* network of $|V|$ nodes is: $\frac{|V-S|}{|S|}$. The *expansion quality* $\frac{|N(S)|}{|V-S|}$, then, captures the extent to which a sample achieves this maximum possible expansion. A score of 1 indicates that the sample "touches" or is one hop away from every other node in the network.

[3]The *expansion* of an entire graph is typically taken to mean $\min_{S \subset V} \frac{|N(S)|}{|S|}$ [12].

[4]Alternatively, one can normalize *expansion* as $\frac{|N(S) \cup S|}{|V|}$.

3.2 Datasets

We study expansion and search in a total of ten different networks: two random graph models, a neural network, a power grid, a co-authorship network, an email network, a citation network, a P2P file-sharing network, and two online social networks. It should be noted that not all of these networks may require efficient decentralized search (e.g. a co-authorship network, the neural network of a worm). Nevertheless, these datasets represent a rich set of diverse networks from different domains. This allows us to more comprehensively study network expansion and thoroughly assess the performance of decentralized search strategies in the face of varying network topologies. Table 1 shows characteristics of each network. We now describe each dataset.

Erdos-Renyi Model. One of the first random graph models proposed was that of Erdos and Renyi [6]. The Erdos-Renyi $G(n, p)$ model produces a random graph of n nodes with each of the $\binom{n}{2}$ possible edges existing with probability p. Erdos-Renyi graphs exhibit the short average path lengths found in many real-world networks, but lack the high clustering and skewed (or heavy-tailed) degree distributions found in reality.

Barabasi-Albert Model. The Barabasi-Albert model follows a more, realistic generative process than previous models: the preferential attachment model [2]. A graph of n nodes is grown in a sequential fashion. Each subsequent node of m edges is preferentially attached to previously added nodes with high degree (where the "degree" of a node is the number of neighbors). Graphs generated by this model exhibit skewed, power law degree distributions and short average path lengths, but lack the high clustering found in real networks. (Skewed degree distributions are ones in which there are many nodes with low connectivity and a few nodes with high connectivity that act as hubs. A power law distribution is one such example.)

C. elegans Neural Network is the neural network of the C. elegans worm [33].

Power Grid. This technological network represents the power grid of the western United States [33].

CondMat. This is a co-authorship network of scientists publishing in Arxiv Cond-Mat (i.e. the Condensed Matter Physics category) from the e-print archive, arxiv.org [20].

Enron Emails is the network comprised of email communications among Enron employees [19].

HEPPh is a citation network between papers in Arxiv HEP-Ph (high energy physics phenomenology) from the e-print archive, arxiv.org [8, 20].

Gnutella. This network is an August 31st, 2002 snapshot of the Gnutella peer-to-peer file-sharing network. Nodes represent hosts and edges represent connections among the hosts [29].

Epinions is a trust-based online social network of the consumer review site, Epinions.com [28].

Slashdot is an online social network of the technology news site, Slashdot.com [21].

4. EXPANSION SIGNATURES

In this section we introduce the concept of *expansion signatures*, which concisely captures the expansion properties of a network at different size scales. Intuitively, the *expansion signature* plots the maximum (and minimum) *expansion qualities* of samples at increasing sample sizes. As dis-

Random Graphs	N	D	PL	CC	AD
Erdos-Renyi	10,000	0.0005	4.2	0.0005	6.0
Barabasi-Albert	10,000	0.0005	3.0	0.006	6.0
Real-World	N	D	PL	CC	AD
C. elegans	297	0.05	2.5	0.3	14.5
Power Grid	4941	0.0005	19	0.11	2.7
CondMat	21,363	0.0004	5.4	0.70	8.5
Enron	33,696	0.0003	4.0	0.71	10.7
HEPPh	34,401	0.0007	4.3	0.30	24.5
Gnutella	62,561	0.00008	5.9	0.01	4.7
Epinions	75,877	0.0001	4.3	0.26	10.7
Slashdot	82,168	0.0001	4.1	0.10	12.2

Table 1: Network Properties. **Key:** *N= # of nodes, D= density, PL = characteristic path length, CC = clustering coefficient, AD = average degree.*

cussed, we are mostly interested in samples with the *maximum* expansion, but we include the *minimum* expansion for completeness. As we will see later, *expansion signatures* reveal a number of interesting aspects of networks. But first, we address how precisely to compute the *expansion signature* of a network.

4.1 Problem Formalization

To construct the *expansion signature*, we must seek out the sample S of size k with the maximal (and minimal) expansion for progressively increasing values of k. In Definition 5, we defined the *maximum expander set* as the sample of size k with the maximum expansion. We now formally define the problem of finding this sample:

Definition 7. (*Maximum Expansion Problem*) Given a graph $G = (V, E)$ and a sample size $k < |V|$, the MAXIMUM EXPANSION problem (**MEP)** is to find a sample $S \subset V$ of size k with the maximum expansion, $\frac{|N(S)|}{|S|}$. That is, find: $\text{argmax}_{S:\,|S|=k} \frac{|N(S)|}{|S|}$.

The hardness of various problems related to expansion is well-studied (e.g. [12]). For instance, determining that a graph is a γ-expander (where every $S \subset V$ has expansion of at least γ and $|S| < \frac{|V|}{2}$) is known to be co-NP-complete [12]. The DOMINATING SET problem [10], known to be NP-hard, is to find the smallest sample S such that $N(S) = V - S$. MAXIMUM EXPANSION is clearly a generalization of DOMINATING SET and is, thus, also NP-hard. There also exist reductions to and from the MAXIMUM COVERAGE problem (defined in [10] and also below). Proposition 1, for instance, shows NP-hardness by reduction from MAXIMUM COVERAGE.

PROPOSITION 1. *The* MAXIMUM EXPANSION *problem is NP-hard.*

PROOF. We show a reduction from the MAXIMUM COVERAGE problem, which is known to be NP-hard [10]. In MAXIMUM COVERAGE, given a set $\mathcal{U}$ of n elements, a collection $\mathcal{F} = \{C_i \mid i \in I\}$ of $|I|$ subsets of $\mathcal{U}$ where $\bigcup_i C_i = \mathcal{U}$, and an integer $k < |I|$, the goal is to find k subsets of $\mathcal{F}$ such that their union has the maximum cardinality. To construct a graph G for the MAXIMUM EXPANSION instance, for each $i \in I$ we create a node i. For each element u in $\mathcal{U}$, we create a node u. Thus, $V = I \cup \mathcal{U}$ is the vertex set of G. To create the edge set E of G, an edge $\{i, j\}$ is created for each pair $i, j \in I$ (forming a clique among nodes in I). In addition, for each $i \in I$ and $u \in C_i$, an edge $\{i, u\}$ is created (forming an independent set among nodes in $\mathcal{U}$).

If $C = \{C_i \mid i \in S\}$ is a feasible solution to the MAXIMUM COVERAGE instance for some subset $S \subset I$ where $|S| = |C| = k$, then S is a sample of nodes in G with maximum expansion where $|S| = k$. By construction, each set $C_i \in C$ (where $i \in S$) is represented by a node $i \in I$ from G that is both connected to every other node in I and connected to the nodes in $\mathcal{U}$ that represent elements *contained* by C_i. Thus, S is a k-size sample with the largest neighborhood in G. Conversely, let $S \subset V$ be a sample in G with the maximum expansion. Note that, if $S \cap \mathcal{U} \neq \emptyset$, then a new sample S', with $\frac{|N(S')|}{|S'|} \geq \frac{|N(S)|}{|S|}$, can be constructed by replacing each node $v \in S \cap \mathcal{U}$ with one of v's neighbors $w \in I$. Thus, $C = \{C_i \mid i \in S'\}$ is a feasible solution to MAXIMUM COVERAGE, since each node in S' represents a set in $\mathcal{F}$.

□

Given the hardness of expansion-related problems, one typically resorts to spectral analysis, as the spectrum of a graph can be computed in polynomial time [12]. A key difference in our work, however, is that we are not only interested in the magnitude of expansion, but the *identity* of the sample producing it. Moreover, we are most interested in the *maximum* expansion (as opposed to the minimum expansion, which is normally the focus in theoretical work on expander graphs). Our ultimate objective is to access these high expansion nodes during the course of a decentralized search to understand and facilitate search performance.[5] Spectral analysis may be less useful here. Thus, we approximate expansion using a simple greedy algorithm (GREEDYAPX). At each iteration, we greedily select the node that maximizes (or minimizes) the expansion of the currently constructed sample, as shown in Algorithm 1. We now show that this simple greedy algorithm yields a $(1 - 1/e)$-approximation guarantee for the MAXIMUM EXPANSION problem.

PROPOSITION 2. GREEDYAPX *approximates* MAXIMUM EXPANSION *within a ratio of at least* $1 - 1/e \approxeq 0.632$.

PROOF. The structure of this proof follows that of the well-known proof for the MAXIMUM COVERAGE greedy approximation (see [7, 11], for instance). Let S_{opt} be the optimal sample of size k and N be the set of nodes covered by S_{opt} (where "covered" is taken to mean $N(S_{opt}) \cup S_{opt}$). Let N_i be the set of *new* nodes covered by the i^{th} iteration of GREEDYAPX. Since N can be covered by a sample of size k, by the pigeonhole principle:

$$|N_i| \geq \frac{|N| - \sum_{j=1}^{i-1} |N_j|}{k}$$

[5]For directed networks that are very weakly connected, a sample with high maximum expansion (based on out-degree) may exist, but the nodes in the sample itself may not be reachable from substantial portions of the network. Samples such as this may shed little light on searchability. One possible approach to address these scenarios is to compute expansion signatures using the *expected* maximum expansion of *connected* samples. In the present work, however, for simplicity and brevity, we treat all links as bidirectional (or undirected).

Algorithm 1 GREEDYAPX

Input:
  Graph $G = (V, E)$
  k, the sample size.
$S = \emptyset$ // initialize sample to empty set
$v = \text{argmax}_w |N(\{w\})|$
$S = S \cup \{v\}$
while $|S| \leq k$ **do**
  Select new node $v \in V - S$ that maximizes (or minimizes):
    $|N(\{v\}) - (N(S) \cup S)|$
  $S = S \cup \{v\}$
end while

Then, $\sum_{j=1}^{i} |N_j| \geq |N| - |N|(1 - \frac{1}{k})^i$ and

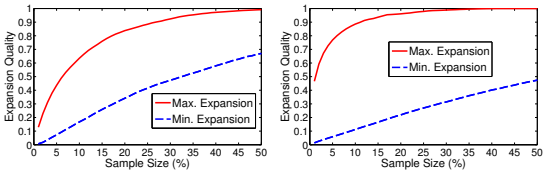
$$\sum_{i=1}^{k} |N_i| \geq |N| - |N|(1 - \frac{1}{k})^k \geq |N|(1 - \frac{1}{e}).$$

□

It should be noted that, during preliminary testing, we also experimented with using simulated annealing for finding the sample with maximum expansion, but GREEDYAPX was shown to be superior. We now use GREEDYAPX to construct *expansion signatures* for both synthetic random graphs and real-world networks. We discuss each separately.

4.2 Signatures for Random Graphs

We examine the *expansion signatures* of two well-known random graph models: Erdos-Renyi (ER) graphs [6] and the Barabasi-Albert preferential attachment model (BA) [2]. The ER and BA models produce graphs with very different degree distributions. Whereas the BA model produces graphs with highly skewed, heavy-tailed degree distributions that follow the power law [2], the ER model produces graphs following a Poisson degree distribution [6]. It is clear that the BA model exhibits a higher and more rapidly increasing maximum expansion (which is a result of the highly connected hubs in its skewed degree distribution). However, the ER model exhibits a relatively higher *minimum* expansion. In fact, random d-regular graphs, where every node has the same degree d, also have "good" minimum expansion with high probability (where *every* $S \subset V$ will have high expansion) [12]. In Section 6, we will examine whether it is the maximum or minimum expansion that most affects searchability.



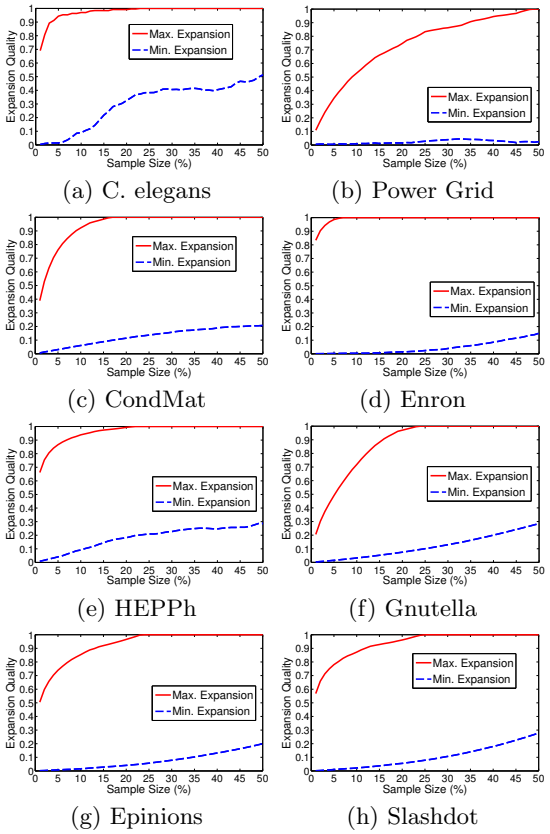
(a) Erdos-Renyi (b) Barabasi-Albert

Figure 1: *Expansion Signatures* for ER and BA models.

4.3 Signatures for Real-World Networks

We now turn our attention to the *expansion signatures* of real-world networks. We examine eight different networks from diverse domains. *Expansion Signatures* for each network are shown in Figure 2. We can immediately see that different types of networks exhibit very different expansion properties. For instance, the size scale required to obtain a maximum *expansion quality* of 1 in the Enron network is only 7%. For the power grid, it is 49%.

We also see that the *minimum* expansion varies across networks. We identify two different causes for low minimum expansion: 1) there is *extreme* sparsity in the number of edges (imagine a simple line graph) or 2) the network is relatively sparse while exhibiting a high degree of clustering (imagine small sets of densely linked nodes linked together by *sparse* connections). Both cases result in a relatively low minimum expansion (as the neighborhood size ($|N(S)|$) will be small for most samples). In the former case, the maximum expansion will tend to also be low (e.g. Power Grid). In the latter case, we find the maximum expansion to be relatively higher (e.g. CondMat, Enron). As will be discussed later, we posit the sparse links between relatively dense clusters in networks result in these higher values for maximum expansion. Recall also that the minimum expansion is related to the classic definition of an expander graph: every S has expansion at least γ in a γ-expander [12]. The co-authorship, email, and social networks, with higher clustering and consequent lower minimum expansion, do *not*, then, appear to be classic expander graphs. In Section 6, we will see whether or not this low minimum expansion affects searchability and information dissemination.



(a) C. elegans (b) Power Grid
(c) CondMat (d) Enron
(e) HEPPh (f) Gnutella
(g) Epinions (h) Slashdot

Figure 2: *Expansion Signatures* for different networks.

5. SEARCHING NETWORKS

We now address the problem of decentralized search in networks. In a typical realistic scenario, starting from some initial node, we must locate some other node in the network *without* full knowledge of global network topology. Thus, we are unable to simply compute the shortest path, and we must hop from node to node until the destination node is found. The running example application we employ is querying unstructured peer-to-peer file-sharing networks, where a search is comprised of sending a query message from node to node. The destination node in question, for instance, might host a particular file of interest to the querier. How can we locate this destination node? Flooding the network with the query (where a node receiving a query forwards it to all neighbors) is provably unscalable and impractical [1, 30]. In fact, when the music file-sharing service Napster became unavailable due to a court injunction in 2001, the Gnutella network (which employed a flooding-based search protocol at the time) crashed due to the large influx of former Napster users [30].

In Section 4, through *expansion signatures*, we have seen that it is often a relatively small set of nodes that is connected to a large portion of the network. If one were able to easily locate this set of high expansion nodes, then the efficiency of search might vastly be improved (as this would quickly take us within one hop of many other nodes). But, how can these nodes be accessed during the course of a decentralized search? Our aforementioned greedy $(1 - 1/e)$-approximation algorithm to find high expansion nodes, shown in Algorithm 1, assumes we have access to the network in its entirety (in which case decentralized search would not even be needed). As mentioned, we are interested in cases where there is *limited* knowledge of global network connectivity. Therefore, we adapt the greedy $(1 - 1/e)$-approximation algorithm from Section 4 into a greedy search heuristic - one that does *not* require full knowledge of network topology. We refer to this search heuristic as an *expansion search*. We compare the *expansion search* to several popular search strategies in complex networks. These include a *degree search* [1], a breadth-first search (BFS) [14, 15, 34], and a random walk [1, 22]. We note that, in a BFS-based search strategy, there are *multiple* copies of the search query traversing the network. In contrast, for the remaining three search strategies, there is a *single* copy of the search query. For all search methods, unvisited nodes are always preferentially selected over previously visited nodes at each step in the search. We now describe each search strategy in detail.

Expansion Search (XS). In an *expansion search*, the next node in the search is selected so as to maximize the expansion. Let S be the set of nodes visited thus far, let $N(S)$ be the neighborhood of the visited nodes, and let c be the current, most recently visited node (where $c \in S$). Then, in an *expansion search*, the next hop is selected from among the unvisited neighbors of c (i.e. $N(\{c\}) - S$) as the node that maximizes the expansion. That is, we visit node v where

$$v = \underset{v \in N(\{c\}) - S}{\operatorname{argmax}} |N(\{v\}) - (N(S) \cup S)|$$

The key difference, then, is that the next hop is selected from the neighborhood of the current node c (i.e. $N(\{c\})$), rather than all of $V - S$ (as is the case in the greedy approximation algorithm described in Algorithm 1).

Degree Search (DS). The degree-based search was proposed by Adamic et al. [1]. At each step in the search, the search query is forwarded to the unvisited neighbor with the highest degree (i.e. largest number of neighbors). That is, the next hop selected is node v where

$$v = \underset{v \in N(\{c\}) - S}{\operatorname{argmax}} |N(\{v\})|.$$

Adamic et al. [1] analytically and empirically showed that, for power-law networks, if nodes with highest degree are preferentially selected during the search and visited first, substantial portions of the network can be covered and explored.

Breadth-First Search (BFS). One type of search strategy used most often in practice is a breadth-first search [26, 31]. In its simplest form, this involves flooding the network, where each node sends a copy of the query to each and every one of its neighbors. These flooding and broadcast methods find targets quickly. But, as we have already mentioned, they are highly unscalable due to the tremendous overhead incurred from redundant forwards (as each node forwards the query regardless of whether its neighbors have already received it). As a result, a number of variations on flooding have been proposed to reduce this overhead (e.g. [14, 15, 34]). In this work, we evaluate a hypothetical BFS strategy in which there are *no* redundant messages. In other words, each message holder forwards a copy of the query only to those neighbors who have not yet received it, and all copies of the query terminate as soon as at least one copy of the query is successful and reaches its destination. Note that this avoidance of redundant forwards and immediate termination are somewhat unrealistic for BFS or flooding strategies. Unlike the other search methods we evaluate, there are *multiple* copies of the query traversing the network in a BFS-based strategy. And, with no information transfer between the various copies of the query, it is difficult to determine which neighbors have already seen the query or when one of the copies reaches the intended target. Nevertheless, our implementation of pure BFS allows us to test the *true* power of flooding-based strategies. If this strategy, with its unrealistic and unfair advantage, still cannot match the performance of other search strategies, then BFS-based methods may not hold as much promise as previously thought, and their utility for exploring networks (e.g. P2P, focused web crawling) should be possibly re-assessed.

Random Walk (RW). The final search strategy we evaluate is the random walk [1, 22] in which the current node forwards the query to exactly one randomly selected neighbor. We employ a *self-avoiding* random walk [1] where the next hop is selected randomly from among the neighbors who have not yet been visited in the search. Note that, as opposed to BFS-based strategies, self-avoidance to eliminate redundant forwards is realistic here because there is a *single* copy of the query traversing the network, within which a list of previously visited nodes can be stored. (The same is true for self-avoidance in the *expansion search* and the *degree search*.)

We conclude this section with two final remarks. First, for both the *expansion search* and *degree search*, each node must

know both its neighbors *and* its neighbors' neighbors. This is required so that the *expansion search* and *degree search* can compute expansion and degree (respectively). This, as it happens, is a modest and satisfiable requirement for many application domains. For instance, in a P2P network, nodes must communicate with their neighbors when joining or leaving the network anyway and neighbor lists can be exchanged during this communication. In fact, several existing search protocols exchange information with nodes at distances of even greater than two hops [1, 31]. Even in a social network, one typically is aware of friends of friends.

Second, as mentioned previously, unvisited nodes are always preferentially chosen over visited nodes in all four search strategies. But, at some points during the search, it may be the case that all the neighbors of a given node are already visited. There are several approaches to dealing with these situations. The next step might be chosen uniformly at random from among the visited neighbors, for instance. For the *expansion search* and the *degree search*, another approach is to select the "best" unvisited node from among the neighborhood of all previously visited nodes (i.e. if S is the set of visited nodes, select a node $v \in N(S)$ with the highest degree or best expansion). Note that, if using this approach, the partial topology of the network, learned during the course of the search, must be stored so that a path to the best next hop may be traversed. During preliminary testing, we did not find a significant performance difference between the two. Therefore, we only consider the former approach: when all neighbors of a current node are visited, the next hop is selected uniformly at random from among these visited neighbors.

6. EXPERIMENTAL EVALUATION

6.1 Experimental Setup

We evaluate each search strategy on each network and track performance over time. Each node is assumed to know its neighbors and passes received messages to them based on one of the four search strategies. The search ends when the message is passed to a neighbor of the target, at which time the message-holder can pass the message directly to its destination[6]. We track the cumulative nodes discovered[7] at each step of a search, which is comprised of both the nodes visited and the neighbors of nodes visited. We define a "step" in the search as a single hop taken by a single query message. If there are multiple copies of the message (as in the case of a BFS or flooding strategy), then the number of steps is defined as the total number of hops taken by all copies of the message in the system. Note that this setup is somewhat of a worst case scenario, as we are assuming there is but a single node in the entire network capable of satisfying a given search query. In the context of a P2P network, for instance, we are assuming that there is a single file residing on a single node in the whole network that must be located. As a result, actual performance in real applications, where multiple nodes can satisfy a search query, will be much higher. The extent will be domain-specific and depend on the extent of object (or file) replication in the network. This setup, then, allows us to evaluate the performance of each search strategy *independent* of the effects of extraneous factors such as replication.

[6]In the context of P2P, we assume each node knows the *identity* of its neighbors' neighbors, but not necessarily the *files* stored by its neighbors' neighbors.

[7]For the Experimental Evaluation section, we employ the normalized cumulative nodes discovered ($\frac{|N(S)\cup S|}{|V|}$) as the evaluation measure rather than the *expansion quality* ($\frac{|N(S)|}{|V-S|}$).

6.2 Experimental Results

6.2.1 On the Performance of Search Strategies

We first examine the relative performance of each search strategy on each network. Table 2 shows the number of steps required to discover 20%, 35%, and 50% of the nodes in the network. As can be seen, the *expansion search* (XS) exhibits the best overall performance. We also find a clear performance difference between the conventional search strategies (BFS and RW) and the less conventional approaches (XS and DS). We discuss each separately.

XS and DS Performance

Overall, we find the XS and DS strategies to exhibit the best general performance with the XS approach faring better. On most of the networks, the XS strategy either exceeds or ties the performance of other approaches. This leads us to a natural question: what causes performance differences between XS and DS? On networks in which XS and DS perform similarly, high degree nodes will tend to link to different sets of nodes (in which case high degree nodes and high expansion nodes will tend to be one and the same and will discover a similar amount of nodes). On the other hand, for networks where XS exceeds the performance of DS, we posit that these nodes may be more likely to have similar neighbors, in which case a high degree node may, in fact, have *low expansion* if it links to neighbors already seen during the search. In these cases, the XS strategy will discover more nodes. Overall, despite the modestly better performance of the XS method, we find the DS strategy performs exceedingly well, which indicates that the former case may be more common in real-world networks. That is, on real-world networks, a *degree search* may do well in finding high expansion nodes without explicitly looking.

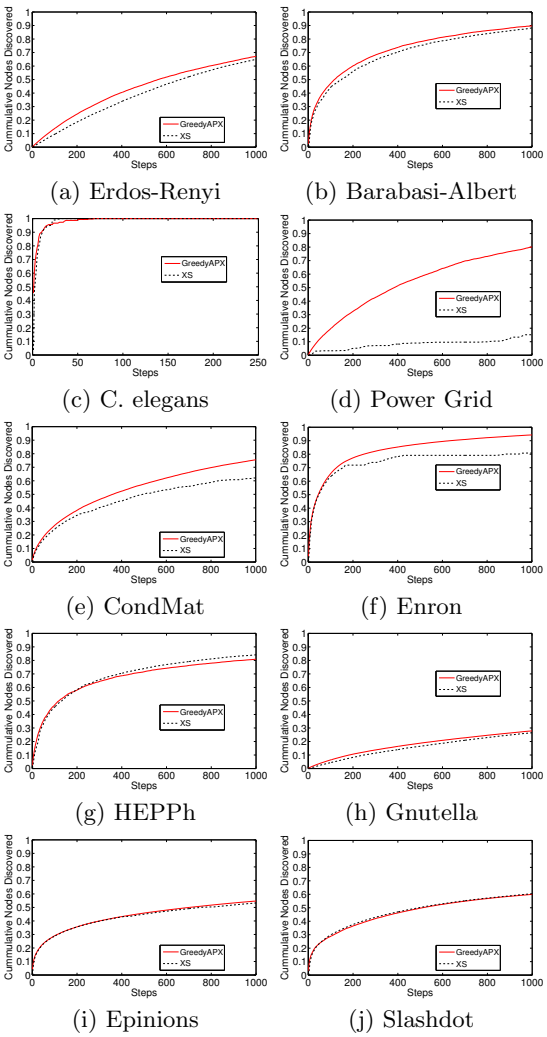
The one network on which neither XS nor DS performs the best is the power grid. The power grid seems to be the least well-connected network evaluated (with mean degree of only 2.7). In fact, it has such low connectivity that only a systematic BFS does best in exploring the network.

BFS and RW Performance

Conventional approaches to searching networks such as P2P systems include those based on random walks (RW) and breadth-first search (BFS) [26]. As mentioned, flooding strategies, based on BFS, are used most often in real applications, as they tend to find answers quickly. BFS is also pervasively used in web crawling and graph sampling. It is striking, then, that our idealized version of BFS, one that avoids redundant communications and immediately terminates upon success, still cannot outperform other approaches (on all but the power grid). In general, we find that the BFS and RW approaches exhibit a relatively lower *expansion quality* as compared to XS, fail to explore the network as well as other strategies in the same number of message forwards, and, consequently, discover less nodes.

Comparison to GreedyAPX

Given the relatively better performance of the XS strategy as compared to other methods, we now examine the extent to which it matches the performance of GreedyAPX (our best known approximation of the maximum expansion in a network). Figure 3 shows the cumulative nodes discovered by GreedyAPX and XS for the first 1000 steps. Interestingly, the XS strategy, which hops from node to node and performs the search *without* complete access to the network in its entirety, often comes close to the performance of GreedyAPX (which *does* have random access to the entire network). Once again, the most salient exception is the power grid, which seems to be the least searchable network evaluated. We discuss network searchability next.



(a) Erdos-Renyi (b) Barabasi-Albert

(c) C. elegans (d) Power Grid

(e) CondMat (f) Enron

(g) HEPPh (h) Gnutella

(i) Epinions (j) Slashdot

Figure 3: [**Best viewed in color.**] Comparison of GreedyAPX and XS for first 1000 steps of a search. In most cases (save for the power grid), XS strategy closely matches GreedyAPX (our best approximation for the maximum expansion).

6.2.2 *On the Searchability of Networks*

Expansion and Searchability

From Figures 1 and 2 and Table 2, we can see that the magnitude of maximum expansion (as approximated by GreedyAPX), corresponds remarkably well to the extent to which each network is searchable. On any given network, when the maximum expansion is low, *all* search strategies perform significantly worse. On the other hand, when the maximum expansion is high, *all* search strategies fare relatively better. The *expansion signatures*, then, correctly infer the ease of search and information dissemination in a network.

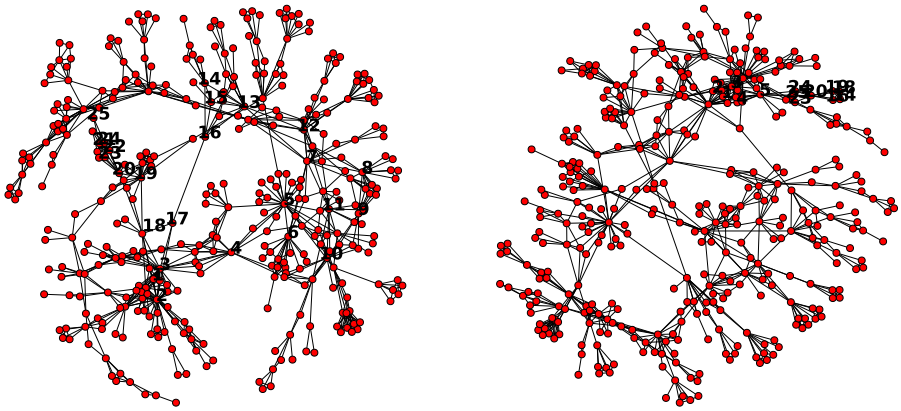
It is also striking to find that it is the *maximum* expansion (rather than the *minimum* expansion) most responsible for the level of searchability in a network. The classic definition of an expander graph is based on *minimum expansion*. Recall that a graph is a γ-expander if $|N(S)| \geq \gamma|S|$ for each $S \subset V$ where $|S| \leq \frac{|V|}{2}$ [12]. In the literature, expander graphs and minimum expansion are often connected to the ease of dissemination in network (e.g. [3, 5]). For instance, [3] has claimed social networks to be expander graphs as a means to explain the ease of diffusion across them. In contrast, our work shows that social networks are *not* classic expander graphs and have a low *minimum* expansion due to clustering. Moreover, we find that it is the *maximum* expansion, not the minimum expansion, that is related to efficient searchability in social networks and other graphs.

Structural Properties and Possible Explanations

It is both surprising and ironic that the Gnutella network, which exists for the very purpose of search, turns out to be one of the *least* searchable networks we evaluated. The only other network exhibiting *less* searchability is the power grid. At the other end of the spectrum, the C. elegans and Enron networks appear to be the *most* searchable. As shown in Table 2, for both Enron and C. elegans, all four search strategies are able to discover half the network in a very small number of steps. What causes a network to exhibit high maximum expansion and good searchability?

One of the more obvious explanations is that denser, more well-connected networks tend to be more searchable than extremely sparse networks. Nodes have larger neighborhoods in denser networks and are, therefore, easier to explore. This is true for the same reason a clique is intuitively more searchable than a long sequence or chain of nodes each connected by a single edge. For instance, the "unsearchable" power grid has a density of 0.0005 and mean degree of only 2.7 whereas the C. elegans network has a density of 0.05 and mean degree of 14.5. The Gnutella network, like the power grid, is also relatively more sparse than other networks of equivalent size.

Density, however, fails to explain the whole story. Consider the ER and BA graphs. Both were constructed to have similar densities but exhibit different *expansion signatures* (see Figure 1) and, correspondingly, different degrees of searchability (see Table 2). By virtue of its skewed degree distribution, the BA model seems to exhibit better searchability than that of the ER model, and the effect of degree distributions on search and dissemination is well-known (e.g. [1, 2]). By visiting well-connected hubs, one can quickly cover significant portions of a network. But, once again, degree distributions fall short in adequately explaining the ease of search. Many of the networks considered exhibit skewed, heavy-tailed degree distributions (e.g. Enron, Epinions), but, nonetheless, exhibit different levels of searchability. Surprisingly, in stark contrast to previously



(a) Expansion Search (XS) (b) Random Walk (RW)

Figure 4: Numbers on each plot show the trace of the first 25 steps in a search by an *expansion search* (XS) and a self-avoiding random walk (RW). Both searches were started from the same initial source node. The XS strategy explores a wider portion of the network and more clusters in the same number of steps

.

held beliefs (e.g. [13]), even average path length fails to fully explain searchability. A number of networks have very similar average path lengths (see Table 1), but very different levels of searchability (see Table 2).

Unlike density and degree distributions, the effect of clustering on searchability and dissemination is less studied and more nebulous. As mentioned in Section 6.2.1, based on our results, we reason that clustering can also facilitate searchability. Real-world networks often exhibit what is known as *community structure* [9, 32]. Intuitively, a community in a network is a cluster of nodes more densely connected to each other than other nodes and exhibit higher clustering coefficients than one would expect at random [9, 32]. By this intuitive definition, nodes in the same community will be expected to share more neighbors than nodes in different communities (by virtue of the dense connections within clusters and lower conductance). As a result, if one were to visit a small number of nodes from many different communities, the expansion (and, therefore, conductance) of these visited nodes would be high and many nodes would be discovered in the search. By searching based on expansion, more communities (and, consequently, larger portions of the network) are explored[8], and this can be demonstrated. Consider the network theory co-authorship network [27], a small, sparse network considered by many to exhibit some degree of community structure. Figure 4 shows a typical path taken by both an *expansion search* (XS) and a random walk (RW) on this network. The XS strategy, by attempting to maximize expansion, jumps across the boundaries between different clusters more easily and is able to explore larger portions of the network. In this way, clustering and community structure, like high density and skewed degree distributions, can facilitate searchability in a network. However, the only common thread and unifying theme that fully and consistently explains searchability across different networks is the singular concept of *expansion*.

7. CONCLUSIONS

We have introduced the concept of *expansion signatures* and have used them to study the effect of expansion on decentralized search in networks. We have shown that it is the magnitude of maximum expansion (rather than minimum expansion) that corresponds to the extent to which a network is efficiently searchable. Moreover, we have shown that traditional graph properties such as average path length and skewed degree distributions fail, by themselves, to fully explain the level of searchability in a network. Finally, we have shown that a search strategy based on maximizing expansion covers the network far better than some typical approaches to decentralized search. For future work, we plan to further investigate the interplay between expansion and various graph-theoretic properties and their effect on dissemination.

[8]This relationship between the maximum expansion and community structure has been demonstrated in [24].

8. REFERENCES

[1] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 64(4):046135+, Sep 2001.

[2] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[3] C. L. Barret, S. G. Eubank, and J. P. Smith. *Fighting Infectious Diseases (Sci. American)*. Rosen Publishing Group, 2007.

[4] M. Boguna, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nature Physics*, 5(1):74–80, November 2008.

[5] F. Chierichetti, S. Lattanzi, and A. Panconesi. Rumour spreading and graph conductance. In *SODA 2010*.

[6] P. Erdös and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[7] U. Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, July 1998.

[8] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 kdd cup. *SIGKDD Explor. Newsl.*, 5(2):149–151, December 2003.

[9] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, June 2002.

	20%				35%				50%			
	XS	DS	RW	BFS	XS	DS	RW	BFS	XS	DS	RW	BFS
ER	**218**	224	366	386	**417**	443	711	738	**662**	720	1141	1188
BA	**18**	**18**	154	91	59	**56**	288	285	149	**145**	537	393
C. eleg.	**2**	**2**	**2**	3	**2**	**2**	4	7	**3**	**3**	9	8
Power	1394	1450	1271	**649**	2220	2370	2794	**1332**	8051	6151	5148	**2091**
CondMat	**72**	93	474	413	**208**	317	1064	1071	**495**	827	2248	2336
Enron	**9**	10	125	266	**20**	22	342	801	**49**	58	559	1941
HEPPh	**26**	37	204	446	**56**	80	469	1366	**132**	250	825	2205
Gnutella	**659**	720	1788	1730	**1577**	1836	3897	3829	**2930**	3615	7191	6875
Epinions	**34**	48	281	590	**189**	344	1059	2679	**752**	1213	3029	5948
Slashdot	**32**	43	241	338	**163**	239	859	1612	**492**	725	1997	4210

Table 2: Number of steps to discover 20%, 35%, and 50% of the network. The best (i.e. lowest) value is in highlighted for each dataset. Overall, XS performs best. The variance for XS and DS was significantly small and standard error is omitted for ease of illustration. (Standard error for RW/BFS was larger, but not so large that either became a candidate for the best or even second-best performer.)

[10] D. S. Hochbaum, editor. *Approximation algorithms for NP-hard problems*. PWS Publishing Co., Boston, MA, USA, 1997.

[11] D. S. Hochbaum and A. Pathria. Analysis of the greedy approach in problems of maximum k-coverage. *Naval Research Logistics (NRL)*, 45(6):615–627, September 1998.

[12] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc*, 43, 2006.

[13] K. Y. K. Hui, J. C. S. Lui, and D. K. Y. Yau. Small-world overlay p2p networks: construction, management and handling of dynamic flash crowds. *Comput. Netw.*, 50(15):2727–2746, 2006.

[14] S. Jiang, L. Guo, X. Zhang, and H. Wang. Lightflood: Minimizing redundant messages and maximizing scope of peer-to-peer search. *IEEE TPDS*, 19(5):601–614, 2008.

[15] S. Jin and H. Jiang. Novel approaches to efficient flooding search in peer-to-peer networks. *Computer Networks*, 51(10):2818–2832, July 2007.

[16] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *STOC 2000*.

[17] J. Kleinberg. Complex networks and decentralized search algorithms. In *ICM*, 2006.

[18] J. Kleinfeld. Could it be a big world after all? the 'six degrees of separation' myth. *Society*, April 2002.

[19] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *ECML 2004*.

[20] J. Leskovec. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05*.

[21] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08*.

[22] X. Li and J. Wu. Searching techniques in peer-to-peer networks. In *Handbook of Theoretical and Algorithmic Aspects of Ad Hoc, Sensor, and Peer-to-Peer Networks*, pages 613–642, 2006.

[23] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, August 2005.

[24] A. S. Maiya and T. Y. Berger-Wolf. Sampling community structure. In *WWW '10*.

[25] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[26] B. Mitra. Technological networks. In N. Ganguly, A. Deutsch, and A. Mukherjee, editors, *Dynamics On and Of Complex Networks*, chapter 15, pages 253–274. Birkhäuser Boston, Boston, 2009.

[27] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104+, 2006.

[28] M. Richardson, R. Agrawal, and P. Domingos. *Trust Management for the Semantic Web*, volume 2870. January 2003.

[29] M. Ripeanu, I. Foster, and A. Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. Sep 2002.

[30] S. Schmid and R. Wattenhofer. Structuring unstructured peer-to-peer networks. In S. Aluru, M. Parashar, R. Badrinath, and V. K. Prasanna, editors, *HiPC 2007*, volume 4873 of *Lecture Notes in Computer Science*, chapter 40, pages 432–442. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[31] D. Tsoumakos and N. Roussopoulos. Analysis and comparison of p2p search methods. In *InfoScale '06*.

[32] S. Wasserman and K. Faust. *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)*. Cambridge University Press, February 2005.

[33] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[34] B. Yang and H. G. Molina. Improving search in peer-to-peer networks. In *ICDCS '02*.