

Decentralized Search in Networks Using Homophily and Degree Disparity

Özgür Şimşek and David Jensen

Computer Science Department
University of Massachusetts
Amherst, MA 01003-9264
{ozgur, jensen}@cs.umass.edu

Abstract

We propose a new algorithm for finding a target node in a network whose topology is known only locally. We formulate this task as a problem of decision making under uncertainty and use the statistical properties of the graph to guide this decision. This formulation uses the homophily and degree structure of the network simultaneously, differentiating our algorithm from those previously proposed in the literature. Because homophily and degree disparity are characteristics frequently observed in real-world networks, the algorithm we propose is applicable to a wide variety of networks, including two families that have received much recent attention: small-world and scale-free networks.

1 Introduction

In a well known study, Travers & Milgram [1969] asked individuals in Boston, Massachusetts and Omaha, Nebraska to deliver a letter to a target person in Boston, using an unconventional method: the letters were to reach the target person through a chain of acquaintances. The person starting the chain and all subsequent recipients of the letter were given some basic information about the target—including name, address, and occupation—and were asked to forward the letter to someone they knew on a first name basis, in an effort to deliver the letter to the target person with as few intermediaries as possible. Of the 296 letters that were distributed, 22% reached the target, with a median chain length of six.

These findings revealed two surprising properties of the social network—that short paths exist between seemingly unconnected individuals and that people are able to find them—and raised a number of questions: How do people perform this type of search? What properties of the social network make it searchable? And in their presence, how can we search a network efficiently? In this paper we address this last question and present an algorithm for efficient decentralized search in a class of networks that exhibit the properties of the social network of acquaintances.

At the time, the task faced by the participants in the Travers & Milgram study—searching for a target node in a large network whose topology is known only locally—was highly artificial, designed only to explore the structure of the social

network of acquaintances. Today, it appears naturally in various contexts. For example, a similar task is performed when people and focused crawlers [Diligenti *et al.*, 2000; Chakrabarti *et al.*, 1999] search for information in the World Wide Web by following links. The same is true of search protocols that form the backbone of decentralized peer-to-peer file sharing systems such as Gnutella and Freenet [Clarke *et al.*, 2000] that lack a central server to answer queries.

If decentralized search is to succeed, it is essential that the underlying network possess some form of structure that can guide the search. The acquaintance network has at least two characteristics that create such structure. The first is homophily, the *tendency of like to associate with like*, in other words, the tendency of attributes of connected nodes to be correlated—people tend to be acquainted with other people who live in the same geographical area or who have the same occupation. The second characteristic is degree disparity—some people have more acquaintances than others and may act as hubs that connect different social circles. Consideration of homophily gives rise to a message-passing algorithm that favors the neighbor that is the most similar to the target node (e.g., an acquaintance who lives in Boston, if the target person lives in Boston) [Kleinberg, 2000a; 2000b; 2001; Watts *et al.*, 2002], while consideration of degree structure gives rise to an algorithm that favors the neighbor with the highest degree [Adamic *et al.*, 2001].

Building on the insights gained by this recent body of research, we propose a new message-passing algorithm—*expected-value navigation* (EVN)—for decentralized search in networks. Our formulation of the problem is fundamentally different from prior approaches in that it considers the *entirety* of the factors that may influence the effectiveness of the search. We cast the problem faced by each node in the message chain as a decision making task under uncertainty, in which the objective is to minimize the expected length of the search path. This decision is guided by the statistical properties of the graph, in which both homophily and degree play a role. All prior algorithms have essentially used only part of the available information.

Because homophily and degree disparity are characteristics frequently observed in real-world networks, EVN is applicable to a wide variety of networks. We emphasize two families that have received much recent attention: small-world and scale-free networks.

Small-world networks are loosely defined as a family of graphs with a combination of three properties that distinguish them from other graph families (e.g., random graphs, fully connected graphs, and regular graphs): weak connectivity, strong clustering, and small diameter. Many real world networks show the small-world structure, including the World Wide Web, the electrical power grid of the western United States, the collaboration graph of Hollywood actors, and the neural network of the nematode worm *C. elegans* [Watts and Strogatz, 1998]. These networks, by definition, connect most node pairs by short paths, and EVN may be particularly well suited for finding them as the small world structure may arise from homophily [Kleinberg, 2000b].

Scale-free networks are those networks with a power-law degree distribution, which means that the probability of a given degree k is proportional to $k^{-\beta}$, where β is a parameter known as the degree exponent. Most nodes in such networks have only a few edges, but a few nodes have much higher degree. Many small-world networks are scale-free [Barabasi and Albert, 1999].

The remainder of this paper is organized as follows: We first provide a review of the relevant literature. We then present our formulation of the search problem, describe the algorithm we propose, and evaluate it on a collection of synthetic and real-world networks. We conclude with a discussion of our experimental results and directions for future research.

2 Previous Work

A number of message-passing algorithms have been proposed for conducting decentralized search in networks, in which each node receiving the message forwards it to one of its neighbors until the target is found. Based on the decision criteria they use in selecting a forwarding node, these algorithms can be categorized as follows:

- *Degree-Based*—The decision is based on the degree structure of neighboring nodes.
- *Similarity-Based*—The decision is based on how similar the neighboring nodes are to the target node in terms of attribute values.

2.1 Degree-Based Navigation

Adamic *et al.* [2001] proposed an algorithm that forwards the message to the highest-degree neighbor that has not seen the message. On a scale-free network with degree exponent 2.1, in which nodes knew their immediate neighbors and *their* neighbors, this algorithm performed fairly well. Most nodes were easy to find—about 50% of the target nodes were found within 12 hops in a 10,000 node network—but a small proportion of nodes required a much larger hop count. Similar results were obtained on a 700-node subgraph of the Gnutella peer-to-peer file sharing system, which showed a power-law degree distribution with exponent 2.07. The algorithm, however, was not effective on networks with Poisson degree distribution.

2.2 Similarity-Based Navigation

In similarity-based navigation, nodes forward the message to the neighbor that is the most similar to the target node, given a number of attributes on nodes and a similarity metric. This type of search relies on network homophily. Under some conditions, similarity among attribute values of neighboring nodes provides an approximation to a universal gradient that allows short paths to be identified from only local information.

The first algorithmic analysis of similarity-based navigation was performed by Kleinberg [2000b; 2000a] on a simple network motivated by the geographical distribution of acquaintances. This network had nodes on a two-dimensional lattice; each node was connected to all other nodes within a given lattice distance and also to a number of additional nodes across the grid. The probability of a connection of the latter type was proportional to the lattice distance between the nodes raised to the power $-\alpha$, where α is a model parameter called the clustering exponent. In this network, similarity between nodes is defined by their lattice position, and the clustering exponent controls the homophily in the graph: When $\alpha = 0$, long-range contacts are uniformly distributed on the grid; as α increases, long-range contacts become more and more clustered in the node’s vicinity.

Kleinberg showed that when $\alpha = 2$, similarity-based navigation achieves an expected path length bounded by a polylogarithmic function (i.e., a polynomial function of the logarithm) of the number of nodes, and that $\alpha = 2$ is the only clustering exponent at which a polylogarithmic bound on path length is possible. These results generalize to d -dimensional lattices for $d \geq 1$, with the critical value of $\alpha = d$.

Kleinberg [2001] later proved similar results for two other network models that defined node similarity differently, but as in his first model, the link probabilities were a function of the similarity between the nodes and a parameter that controls the degree of homophily in the graph. In both of these models, there was a critical value of the homophily parameter that allowed similarity-based navigation to achieve a search time polylogarithmic in the number of nodes, and for all other values of this parameter, a polylogarithmic upper bound was not possible.

Similarity-based navigation was also explored by Watts *et al.*, [2002], who proposed a hierarchical model of society and a homophily structure that measured similarity with distance in this hierarchy. The authors explored the influence of the number of hierarchies and the homophily parameter on the searchability of the network and found that similarity-based navigation was effective for a large region of the parameter space.

3 Proposed Algorithm: Expected-Value Navigation

We propose a message passing algorithm that builds on the strengths of the algorithms discussed above. We derive this algorithm by formulating the problem as a decision making task under uncertainty, in which the goal is to minimize the expected path length to the target.

The expected value of the path length l_{st} from neighbor s to target t is a weighted sum of all possible path lengths:

$$E(l_{st}) = \sum_{\forall i} i \cdot P(l_{st} = i) \quad (1)$$

We assume that in computing this expected value the following information is available: a list of nodes that have already seen the message, the properties (i.e., degree and attribute values) of neighboring nodes and of the target node, and the known (or estimated) homophily structure of the graph—in other words, a statistical relationship between node similarity and probability of a link. This last piece of information allows us to compute the probability that a given neighbor links to the target node.

We approximate the entire series in Equation 1 using only the first two terms, which are easy to compute given the information available. This estimate captures much of the necessary information because there is no need to know the exact value of the expectation, only whether it is lower than the expectation computed for another neighbor.

If one of the neighbors is the target, this neighbor has $E(l_{st}) = 0$, the lowest possible value. Otherwise, the node for which the second term in the series is the highest minimizes our estimate of $E(l_{st})$ —the larger the probability of a path length of one, the smaller the probability of larger path lengths, and in general, the smaller the expected path length. Note here that for a neighbor that has already seen the message, the second term in the series is zero—we know with certainty that it does not link to the target, otherwise it would have forwarded the message to the target and completed the search.

This gives rise to the following algorithm: If one of the neighbors is the target node, forward the message to this node. Otherwise, forward the message to the unvisited neighbor with the highest probability of having a direct link to the target. If all neighbors have been visited, forward the message to a randomly selected neighbor.

We call this algorithm expected-value navigation (EVN), based on its method of node selection. If the network shows no homophily (i.e., if links are formed independently of node similarity) or if attributes are not available, EVN reduces to the degree-based navigation of Adamic *et al.* [2001]. On the other hand, if degree information is unavailable or if all nodes have equal degree, EVN reduces to similarity-based navigation that avoids visited nodes when possible.

In order to apply EVN in a given network, one needs to compute or estimate the probability that a link exists from one node to another, given the attribute value and degree of both nodes. We estimate this probability assuming that each link is placed independently of the others. For a link from node s to node t , the desired probability p_{st} can then be computed by subtracting from 1.0 the probability that none of the links that originate at s ends at t :

$$p_{st} = 1 - (1 - q_{st})^k \quad (2)$$

where q_{st} is the probability that the first link from s ends at t , and k is the out-degree of node s . This is one of the simplest estimators that uses information on both homophily and

degree—the underlying assumption of independent links is violated in the networks we consider—but our results show that it performs remarkably well.

4 Experimental Evaluation

We evaluated EVN on a collection of synthetic and real-world networks, comparing its performance to three other message passing algorithms: similarity-based, degree-based, and random navigation. These algorithms treat visited neighbors similarly—ignoring them in the presence of unvisited neighbors, and selecting randomly among them otherwise—but differ in how they select among unvisited neighbors. Similarity-based navigation selects the one most similar to the target node in attribute value, degree-based navigation selects the one with the highest degree, and random navigation selects randomly. If more than one neighbor satisfies the criteria, all algorithms select randomly among them.

It is possible to construct other variations of similarity-based, degree-based, and random navigation that differ in how they treat visited neighbors. For instance, the algorithm may ignore prior visitations, or only avoid the last visited node. The versions described above consistently outperformed these variations in our simulations; we therefore do not discuss them any further.

In addition to these four algorithms, we also present the performance of an optimal global algorithm, which returns the shortest path length from source to target if it is less than the number of hops allowed. The performance of the optimal algorithm is a ceiling for the other algorithms—if there is no short path, no algorithm can find it.

We present four performance measures: proportion of successful searches (*prop*), mean path length when successful (*path*), median path length when successful (*median-path*), and mean optimal path length when successful (*opt-path*). The last measure indicates the difficulty of the search tasks an algorithm succeeds at, and is useful when comparing mean path lengths of different algorithms—if one algorithm is able to succeed at more difficult search tasks than another one, the mean path lengths are not directly comparable.

4.1 Synthetic Networks

We considered directed networks with two types of out-degree distribution: power-law and Poisson. We defined a single attribute on each node, which was distributed uniformly in the interval $[0, 1]$. Each network had 1000 nodes; nodes with out-degree higher than 100 were not allowed. Search was terminated after 100 hops if the target was not reached.

The number of outgoing links from each node was determined based on the out-degree distribution of the graph. For a link originating at node s , the probability of linking to node t was proportional to f_{st} , the preference between the two nodes, which we defined as follows:

$$f_{st} = (\max\{|a_s - a_t|, 0.01\})^{-r} \quad (3)$$

where a_s, a_t are attribute values on nodes s and t , and r is a homophily parameter. The *max* term puts a bound on the preference values—in its absence, the preference between two

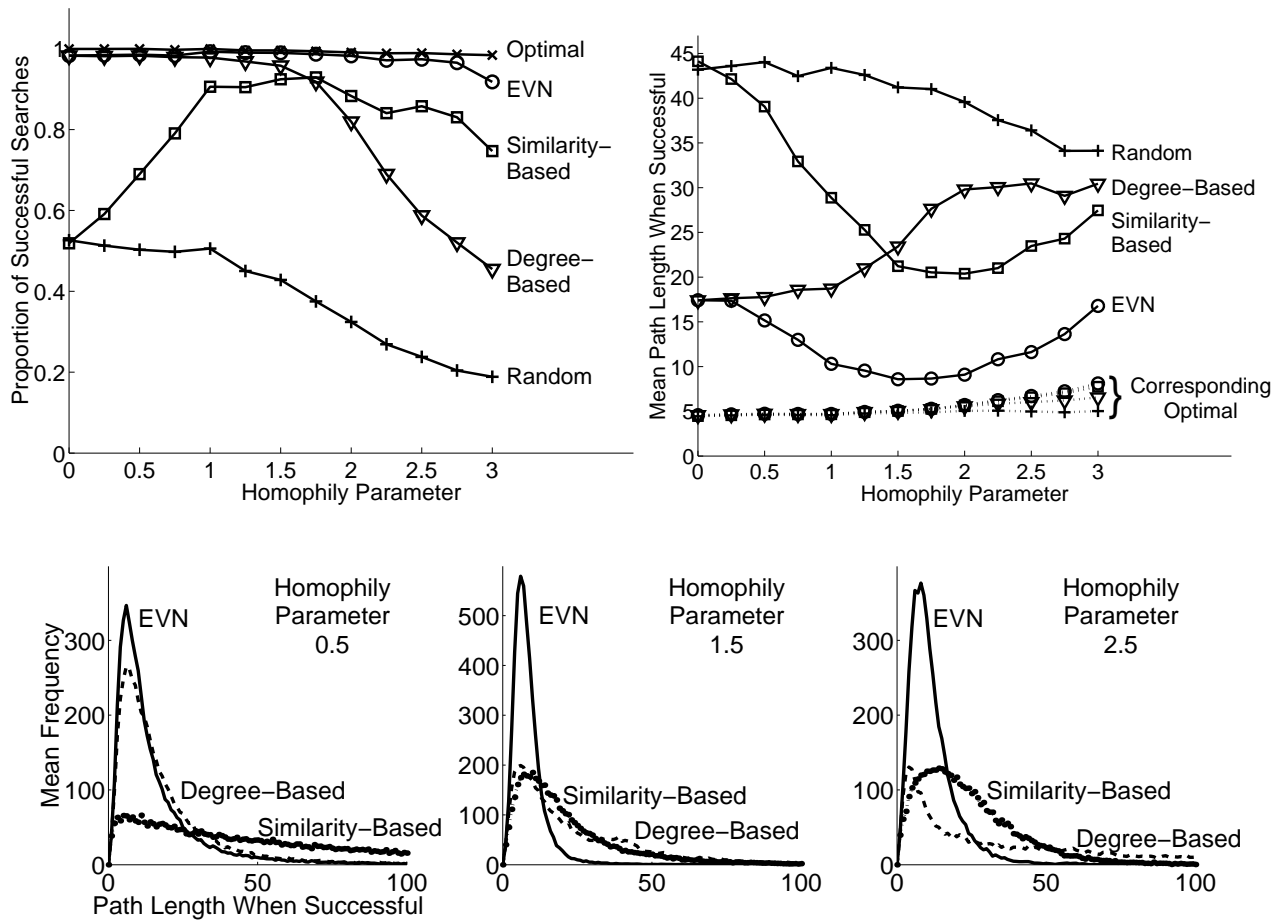


Figure 1: Performance on scale-free networks with degree parameter 1.5.

nodes may be arbitrarily large, as two attribute values may be arbitrarily close.

When r is zero, the graph shows no homophily—a link originating from a given node is equally likely to end at any other node. As r grows, links become more likely to connect nodes with similar attribute values. The findings in Kleinberg [2000b] suggest that small values of r give rise to a homophily structure too weak to guide the search effectively, while large values of r give rise to a graph structure that does not contain short paths.

In applying EVN in these networks, we used $p_{st} = 1 - (1 - f_{st} / (\sum_j f_{sj}))^k$, which was obtained from Equation 2 by substituting q_{st} with $f_{st} / \sum_j f_{sj}$, the ratio of the preference between nodes s and t to the sum of preferences from s to all nodes in the network. In applying similarity-based navigation, we considered all neighbors within 0.01 of the target to be equally close, to account for the presence of the max term in Equation 3.

Networks with Power-Law Degree Distribution

We considered power-law distributions with degree parameters ranging from 1 to 3. This range includes the distributions most frequently observed in real-world networks. The ho-

mophily parameter ranged from 0 to 3. Each possible combination of degree parameter and homophily parameter was evaluated on 10 randomly generated networks (unless noted otherwise), with 5000 randomly selected search tasks in each network.

Figure 1 shows performance on scale-free networks with degree parameter 1.5. In addition to *prop*, *path*, and *opt-path*, this figure also presents the frequency of path lengths when the homophily parameter was 0.5, 1.5, and 2.5. While similarity-based navigation was effective for large values of the homophily parameter and degree-based navigation was effective for lower values, EVN was effective with over 95% success rate for all values of the homophily parameter and returned shorter path lengths than the other algorithms.

Figure 2 shows performance on scale-free networks with varying degree parameters. The data points in this figure show mean values in 30 randomly generated networks. EVN was effective under a wide range of parameter settings and consistently outperformed both degree-based and similarity-based navigation. EVN succeeded at more difficult search tasks than degree-based navigation, as measured by *opt-path*, while returning considerably shorter path lengths. EVN and similarity-based navigation performed similarly in *opt-path*,

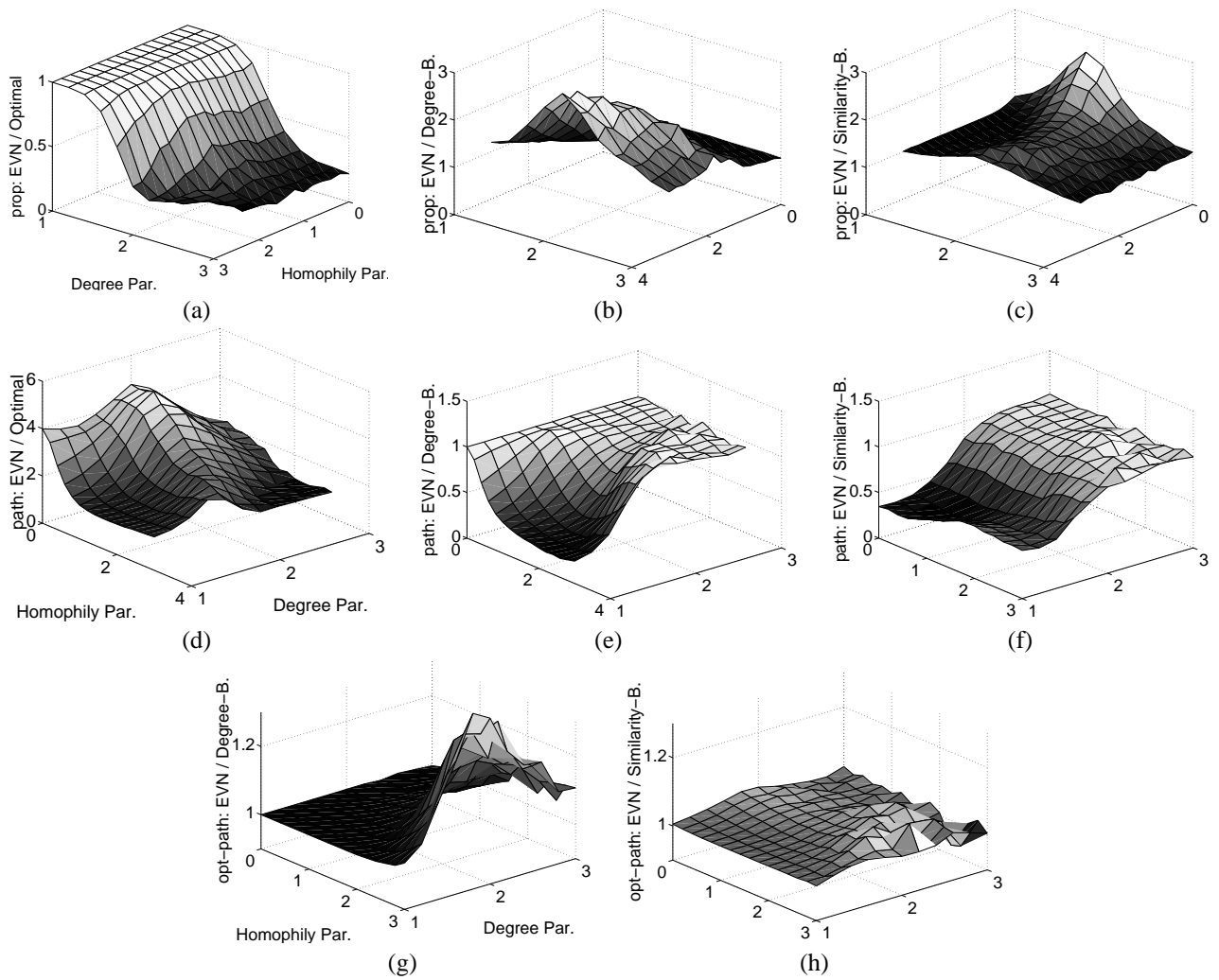


Figure 2: Performance of EVN on scale-free networks. The first row shows the proportion of successful searches by EVN divided by that of (a) optimal, (b) degree-based, and (c) similarity-based navigation; the second row shows mean path length (when successful) for EVN divided by that of (d) optimal, (e) degree-based, and (f) similarity-based navigation; the third row shows mean opt-path of EVN divided by that of (g) degree-based and (h) similarity-based navigation.

but EVN returned shorter path lengths. The proportion of successful searches was higher for EVN than for both degree-based and similarity-based navigation.

Networks with Poisson Degree Distribution

Similar experiments were conducted on networks with a Poisson degree distribution with a mean out-degree of 3.5, which approximately equals the mean degree in the scale-free networks that were tested.

Figure 3 shows performance results. The figure reveals that similarity-based and degree-based navigation were not effective in these networks, succeeding in less than half of the search tasks for all values of the homophily parameter. EVN returned a higher proportion of successful searches than both degree-based and similarity-based navigation; the difference was substantial for a large range of values of the homophily parameter. EVN was most effective with homophily param-

eter values close to 2.

The *path* results shown in Figure 3 may seem counter-intuitive, with random navigation returning the lowest values for most values of the homophily parameter. Recall, however, that *path* refers to the mean path length in *successful* searches, so the mean path lengths are not directly comparable—random navigation returned the lowest *path* values, but it succeed in only the easiest search tasks as measured by *opt-path*. The *opt-path* results show that EVN succeeded at more difficult search tasks than both degree-based and similarity-based navigation. Furthermore, EVN returned considerably shorter path lengths than degree-based navigation, despite succeeding in more difficult search tasks.

Robustness of EVN

In estimating link probabilities, we used the sum of preferences from a given node to all other nodes in the network.

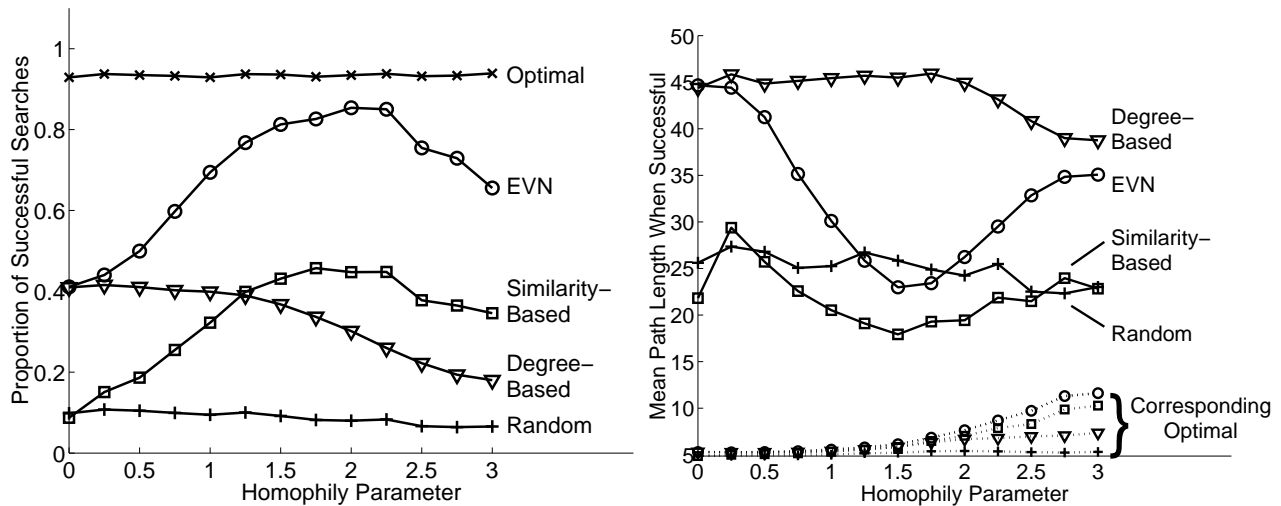


Figure 3: Performance on networks with Poisson degree distribution.

This sum is a normalizing constant that may be thought of as an indicator of network size—it is the sum of similar contributions from all nodes. In experiments we presented so far, we used the exact value of this global constant. Dependence on such global information is clearly not desirable for a decentralized algorithm.

A number of methods for decentralized estimation of global network parameters exist, but all produce local variance in the estimates. Such variance poses little problem for decentralized search algorithms such as EVN, however, because each decision is made locally, and thus estimates of the constant on different nodes need not be consistent. However, serious problems could arise if bias in estimates of the constant degraded the searchability of the network.

We repeated the experiments reported on scale-free networks with degree parameter 1.5 by distorting the normalizing constant with a multiplier of 0.001, 0.01, 0.1, 10, 100, and 1000 for all nodes in the network. Because each node makes its decisions independently of the other nodes, that the constant was distorted in the same way for all nodes is irrelevant. No performance decrements were observed in either performance measure, except when the multiplier was 0.001 or 0.01. Performance decrements for such low values of the multiplier are not unexpected—in these cases, the nodes operated in a 1000 node network, but made decisions as if the network had only a single node or ten nodes. We expect that such poor estimates of the normalizing constant would be avoided easily in practice.

4.2 Scientific Citation Network

We next present results on a real-world network: a citation graph of scientific papers. The nodes in our network were papers from the theoretical high-energy physics (hep-th) area of arXiv.org, an on-line archive of research papers. We included in our network papers that were published in 1995–2000 and had more than 50 non-self citations. The network included 833 nodes and 13,267 links.

Decentralized search in this citation graph is an artificial task—though it does resemble searching for a particular piece of information before the advent of search engines—but the results are useful in evaluating the applicability of EVN to a network that evolves naturally over time, with no known patterns of link formation.

We treated the citation graph as an undirected network, defining node similarity using paper titles and abstracts. The title and abstract of each paper were represented as weighted-term vectors using TFIDF (term frequency \times inverse document frequency) weighting. Paper similarity was computed using a standard cosine correlation measure. We discretized this continuous similarity measure and for each discrete value it took, estimated q_{st} in Equation 2 in a straightforward manner from the network.

We conducted 10,000 randomly selected search tasks, terminating them after 100 hops if the target was not found. Table 1 shows our performance measures; Figure 4 shows the distribution of path lengths returned by each algorithm. Similarity-based navigation was not effective in this task, while degree-based navigation was competitive. EVN performed better than both in all performance measures. Further comparison of EVN and degree-based navigation revealed that EVN succeeded at all search tasks for which degree-based navigation failed and that there were no search tasks in which degree-based navigation succeeded but EVN did not. In those search tasks for which both algorithms succeeded, 47% of the time EVN returned a shorter path than degree-

Algorithm	<i>prop</i>	<i>path</i>	<i>median-path</i>	<i>opt-path</i>
Random	0.89	25.12	20	2.50
Similarity-Based	0.90	23.84	18	2.50
Degree-Based	0.93	9.86	5	2.51
EVN	0.99	6.07	4	2.55
Optimal	1.0	2.55	3	2.55

Table 1: Performance on hep-th citation graph.

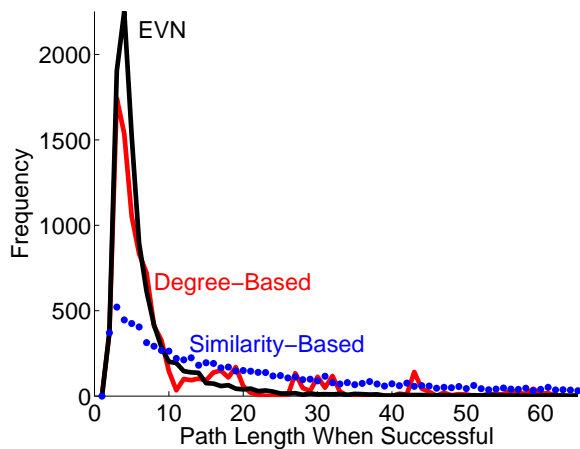


Figure 4: Path length distributions in hep-th citation graph.

based search (mean difference in path length = 11.02); 22% of the time degree-based search yielded a shorter path than EVN (mean difference in path length = 3.75); 31% of the time EVN and degree-based search returned paths of same length.

We next analyzed the sensitivity of the algorithm to the q_{st} values, which were estimated using the entire graph. We distorted q_{st} values (for each discrete value of node similarity) by multiplying them by 0.01, 0.1, 10, and 100. The performance measures were essentially unaffected—they were all within 0.2% of their values reported above.

5 Discussion

We presented a simple and principled algorithm for decentralized search in networks that show homophily and degree disparity. Our formulation of the problem allows one to consider all factors that may influence search performance. In that, it differs fundamentally from previous work in this area, while providing a unifying framework for existing algorithms, which are special cases of the algorithm we present here. Experimental results on a collection of synthetic and real-world networks indicate that our algorithm performs remarkably well, though it may be possible to achieve even better performance by using more sophisticated estimates of the statistical quantities involved.

The utility of our approach depends on the availability of statistical information regarding the relationship between node similarity and link formation. This information is typically available when a network is *designed* by an analyst who establishes the rules on how the network evolves over time (e.g., Zhang et al. [2002]). But we expect that this type of information would also be easy to obtain in other types of networks.

Acknowledgments

We would like to thank Jennifer Neville for providing the data on the hep-th scientific citation graph and to Cynthia

Loiselle for providing comments on earlier drafts of this paper. This research is supported by NSF and Lawrence Livermore National Laboratory (LLNL) under contract numbers HR0011-04-1-0013 and W-7405-ENG-48. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of NSF, LLNL, or the U.S. Government.

References

- [Adamic *et al.*, 2001] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 64, 2001.
- [Barabasi and Albert, 1999] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [Chakrabarti *et al.*, 1999] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific Web resource discovery. In *Proceedings of the Eighth International World Wide Web Conference*, 1999.
- [Clarke *et al.*, 2000] I. Clarke, O. Sandberg, B. Wiley, and T. W. Hong. Freenet: A distributed anonymous information storage and retrieval system. In Hannes Federrath, editor, *Designing Privacy Enhancing Technologies: International Workshop on Design Issues in Anonymity and Unobservability, Lecture Notes in Computer Science 2009*, Berlin, 2000. Springer.
- [Diligenti *et al.*, 2000] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. Focused crawling using context graphs. In *26th International Conference on Very Large Databases, VLDB 2000*, pages 527–534, Cairo, Egypt, 10–14 September 2000.
- [Kleinberg, 2000a] J. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.
- [Kleinberg, 2000b] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.
- [Kleinberg, 2001] J. Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14, 2001.
- [Travers and Milgram, 1969] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.
- [Watts and Strogatz, 1998] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.
- [Watts *et al.*, 2002] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296, 2002.
- [Zhang *et al.*, 2002] H. Zhang, A. Goel, and R. Govindan. Using the small-world model to improve Freenet performance. In *Proceedings of IEEE Infocom*, 2002.