Strong Ties vs. Weak Ties: Studying the Clustering Paradox for Decentralized Search

Weimao Ke
Laboratory of Applied Informatics Research
School of Information and Library Science
& Translational and Clinical Sciences Institute
University of North Carolina at Chapel Hill
wke@unc.edu

Javed Mostafa
Laboratory of Applied Informatics Research
School of Information and Library Science
& Translational and Clinical Sciences Institute
University of North Carolina at Chapel Hill
jm@unc.edu

ABSTRACT

We studied decentralized search in information networks and focused on the impact of network clustering on the findability of relevant information sources. We developed a multi-agent system to simulate peer-to-peer networks, in which peers worked with one another to forward queries to targets containing relevant information, and evaluated the effectiveness, efficiency, and scalability of the decentralized search. Experiments on a network of 181 peers showed that the *RefNet* method based on topical similarity cues outperformed *random walks* and was able to reach relevant peers through short search paths. When the network was extended to a larger community of 5890 peers, however, the advantage of the *RefNet* model was constrained due to noise of many topically irrelevant connections or weak ties.

By applying topical clustering and a *clustering exponent* α to guide network rewiring, we studied the role of *strong ties* vs. *weak ties*, particularly their influence on distributed search. Interestingly, an inflection point was discovered for α, below which performance suffered from many remote connections that disoriented searches and above which performance degraded due to lack of *weak ties* that could move queries quickly from one segment to another. The inflection threshold for the 5890-peer network was $\alpha \approx 3.5$. Further experiments on larger networks of up to 4 million peers demonstrated that clustering optimization is crucial for decentralized search. Although overclustering only moderately degraded search performance on small networks, it led to dramatic loss in search efficiency for large networks. We explain the implication on scalability of distributed systems that rely on clustering for search.

Categories and Subject Descriptors

H.3.4 [**Information storage and retrieval**]: Systems and Software—*Distributed systems, Information networks*

LSDS-IR Workshop. July 2009. Boston, USA.

General Terms

Algorithms, Performance, Experimentation

Keywords

clustering, decentralized search, P2P IR, resource discovery, referral network, agent, weak tie, strong tie, topical locality

1. INTRODUCTION

Information exists in many distributed networked environments, where a centralized repository is hardly possible. In a *peer-to-peer* (P2P) environment, individual peers host separate collections and interact with one another for information sharing and retrieval [18], exemplifying a large, dynamic, and heterogeneous networked information space. Efficient network navigation is critically needed in today's distributed environments, e.g., to route queries to relevant information sources or to deliver information items to peers of interest.

Research has found *clustering* useful for information retrieval. The *Cluster Hypothesis* states that relevant documents are more similar to one another than to non-relevant documents and therefore closely related documents tend to be relevant to the same requests [29]. Traditional IR research utilized document-level clustering to support exploratory searching and to improve retrieval effectiveness [12, 9, 14]. Distributed information retrieval, particularly unstructured peer-to-peer IR, relied on peer-level clustering for better decentralized search efficiency. Topical segmentation based techniques such as semantic overlay networks (SONs) have been widely used for efficient query propagation and high recall [3, 7, 17, 8]. Hence, overall, clustering was often regarded as beneficial whereas the potential *negative* impact of clustering (or over-clustering) on retrieval has rarely been scrutinized.

Research on complex networks indicated that a proper degree of network clustering with some presence of remote connections has to be maintained for efficient searches [15, 25, 30, 16, 24, 6]. Clustering reduces the number of "irrelevant" links and aids in creating topical segments useful for orienting searches. With very strong clustering, however, a network tends to be fragmented into local communities with abundant *strong ties* but few *weak ties* to bridge remote parts [10]. Although searches might be able to move gradually to targets, necessary "hops" become unavailable.

We refer to this phenomenon as the *Clustering Paradox*, in which neither strong clustering nor weak clustering is de-

sirable. In other words, trade-off is required between *strong ties* for search orientation and *weak ties* for efficient traversal. In Granovetter's terms, whereas *strong ties* deal with local connections within small, well-defined groups, *weak ties* capture between-group relations and serve as bridges of social segments [10]. The *Clustering Paradox*, seen in light of strong ties and weak ties, has received attention in complex network research and requires further scrutiny in a decentralized IR context.

In this study, we examined network characteristics and search optimization in a fully decentralized retrieval context. We focused on the effect of network clustering, i.e., strong ties vs. weak ties, on the efficient findability of relevant information sources. Outcome of this research will provide guidance on how an information network can be structured or self-organized to better support efficient discovery of relevant information sources that are highly distributed.

2. RELATED WORK

In an open, dynamic information space such as a peer-to-peer network, people, information, and technologies are all mobile and changing entities. Identifying where relevant collections are for the retrieval of information is essential. Without global information, decentralized methods have to rely on local intelligence of distributed peers to collectively construct paths to desired targets.

2.1 P2P Information Retrieval

In some respect, decentralized IR in networks is concerned with the cost of traversing a network to reach desired information sources. Unstructured or loosely structured peer-to-peer networks represent a connected space self-organized by individuals with local objectives and constraints, exhibiting a topological underpinning on which all can collectively scale [1, 18].

While federated IR research has made advances in enabling searches across hundreds of repositories, a P2P network usually has a much larger number of participants who dynamically join and leave the network, and only offer idle computing resources for sharing and searching [34]. Usually there is no global information about available collections; seldom is there centralized control or a central server for mediating [18, 8].

Recent years have seen growing popularity of peer-to-peer (P2P) networks for large scale information sharing and retrieval [18]. With network topology and placement of content tightly controlled, *structured* peer-to-peer networks have the advantage of search efficiency [27, 21, 5, 19, 26]. However, their ability to handle unreliable peers and a transient population was not sufficiently tested. *Unstructured* overlay systems work in an indeterministic manner and have received increased popularity for being fault tolerant and adaptive to evolving system dynamics [18, 8].

As the peer-to-peer paradigm becomes better recognized for IR research, there have been ongoing discussions on the applicability of existing P2P search models for IR, the efficiency and scalability challenges, and the effectiveness of traditional IR models in such environments [33]. Some researchers applied Distributed Hashing Tables (DHTs) techniques to *structured* P2P environments for distributed retrieval and focused on building an efficient indexing structure over peers [5, 19, 26]. Others, however, questioned the sufficiency of DHTs for dealing with high dimensionality of IR in dynamic P2P environments [3, 18, 17]. For information retrieval based on a large feature space, which often requires frequent updates to cope with a transient population, it is challenging for distributed hashing to work in a traffic- and space-efficient manner.

2.2 Clustering and Decentralized Search

In recent years, topical segmentation based techniques such as semantic overlay networks (*SONs*) have been widely used for P2P IR, in which peers containing similar information formed semantic groups for efficient searches [3, 7, 28, 17, 20]. Clustering, often in the form of hierarchical segments, was the key idea for bringing similar peers together in a more organized way so that topically relevant peers or information sources can be quickly identified. Existing P2P IR research, however, often assumed the unitary benefit of clustering and rarely scrutinized its potential negative impact on decentralized search.

Research on complex networks has found that efficient searching in some properly clustered networks is more promising than in others. Kleinberg (2000) studied decentralized search in small world using a two dimensional model, in which peers had rich connections with immediate neighbors and sparse associations with remote ones [15]. The probability p_r of connecting to a neighbor beyond the immediate neighborhood was proportional to $r^{-\alpha}$, where r was the topical (search) distance between the two and α a constant called *clustering exponent*[1]. It was shown that only when *clustering exponent* $\alpha = 2$, search time (i.e., search path length) was optimal and bounded by $c(\log N)^2$, where N was the network size and c was some constant [15].

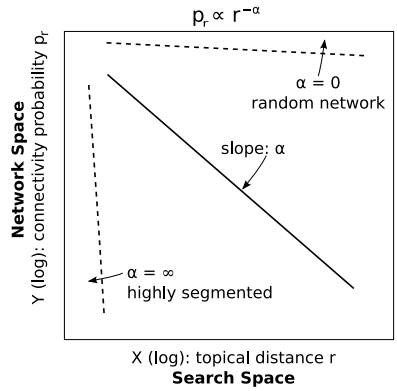


Figure 1: Network Clustering

The *clustering exponent* α, as shown in Figure 1, describes a correlation between the network (topological) space and the search (topical) space [15, 6]. When α is large, weak ties (long-distance connections) are rare and strong ties dominate [10]. The network becomes highly segmented. When α is small, connectivity has little dependence on topical closeness – local segments become less visible as the network is built on increased randomness. In this way, the *clustering exponent* α influences the formation of local clusters and overall network clustering.

It was further demonstrated that optimal value of α for search depends on dimensionality of the search space. Specif-

[1]The *clustering exponent* α is also known as the *homophily exponent* [30, 24].

ically, when $\alpha = d$ on a d-dimension space, decentralized search is optimal. Further studies conducted by various research groups have shown consistent results [30, 16, 24, 6]. These findings require closer scrutiny in an IR context where some assumptions might be voilated, e.g, when orthogonal feature dimensions cannot be precisely defined.

3. APPROACH OVERVIEW

We have developed a decentralized search architecture named *RefNet* for finding distributed information sources in a simulated networked environment. We relied on multi-agent systems to study the problem of decentralized search and focused on the impact of clustering in an information retrieval context. Similar agent-based approaches have been adopted by various research groups to study efficient information retrieval, resource discovery, service location, and expert finding in decentralized peer-to-peer environments [25, 32, 36, 35]. One common goal was to efficiently route a query to a relevant agent or peer[2]. We illustrate the conceptual model in Figure 2 and elaborate on major components.

Assume that agents or peers, representatives of information seekers, providers (sources), and mediators, reside in an n dimensional space. An agent's location in the space represents its information topicality. Therefore, finding relevant sources for an information need is to route the query to agents in the *relevant* topical space. To simplify the discussion, assume all agents can be characterized using a two-dimensional space. Figure 2 visualizes a 2D representation of the conceptual model. Let agent A_u be the one who has an information need whereas agent A_v has the relevant information. The problem becomes how agents in the connected society, without global information, can collectively construct a short path to A_v. In Figure 2, the query traverses a referral chain $A_u \rightarrow A_b \rightarrow A_c \rightarrow A_d \rightarrow A_v$ to reach the target. While agents A_b and A_d help move the query on the horizontal dimension, agent A_c primarily works on the vertical dimension and has a remote connection for the query to jump.

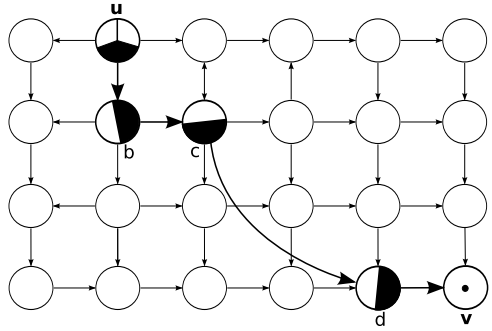


Figure 2: Conceptual Model of RefNet. A circle represents an agent or peer. The black/white segments of each circle illustrate agent representation according to its topical dimensions (coverage).

3.1 Local Indexing & Classification

For decentralized search, direction matters. Pointing to the right direction to the relevant topical space means the agents or peers have some ability to differentiate items on certain dimensions. For instance, one should be able to tell if a query is related to mathematics or not in order to route the query properly on that dimension. Each agent derives clusters or major topics from its local information collection through *document clustering*[3]. The local index provides the basis of an agent's "knowledge" and enables abstraction of queries. Now, when a query is routed to it, the agent will be able to tell what it is about and assign a label to it through *query classification* based on identified clusters [23]. The label associated with the query serves as a clue for potential referral directions.

3.2 Neighbor Selection

Pointing to the right direction also requires that each agent or peer knows which neighbor(s) should be contacted given a labeled query. Therefore, there should be a mechanism of mapping classification output to a potential *good* neighbor. By *good neighbor*, we mean agents on a short path to the targeted information space – either the neighbor is likely to have a relevant information collection to answer the query directly or in a neighborhood closer to relevant targets. Agents explore their neighborhoods through interactions and develop knowledge of who serves or connect to what types of information collections.

3.3 Network Clustering and Rewiring

Network topology plays an important role in decentralized search. Topical segmentation based techniques such as semantic overlay networks (SONs) have been widely used for efficient peer-to-peer information retrieval [8]. Through self-organization, similar peers form topical partitions, which provide some association between the topological (network) space and the topical space to guide searches. Research has found that such an association, in the form of a *clustering exponent* α that defines an inverse relationship between connectivity probability and topical distance, is critical for efficient navigation in networks without global information [15, 16, 6]. The RefNet framework has a mechanism for clustering-based rewiring, which influences the balance of *strong ties* vs. *weak ties* for efficient routing, as illustrated in Figure 1.

4. ALGORITHMIC DETAIL

In the previous section, we proposed and described a conceptual model for decentralized search of relevant information sources. Figure 3 illustrates how various components work together within each agent. This section will elaborate on specific algorithms used in the *RefNet* model for decentralized search.

We used the Vector-Space Model (VSM) for information (document and query) representation [2]. Given that information is highly distributed, a global thesaurus was not assumed. Instead, each agent had to parse information items it individually had and produced a local thesaurus. This thesaurus was then used to represent each information item using the TF*IDF (Term Frequency * Inverse Document

[2]In this paper, the terms *agent* and *peer* are exchangeable.

[3]Note that *document clustering* refers to mining a peer's local collection of documents to identify significant topics and topical overlap whereas *network clustering* is to determine how similar peers connect to each other to form groups and is the main focus of this study.

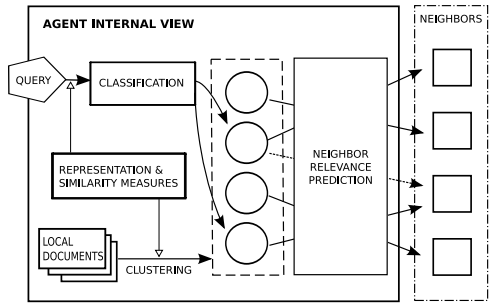


Figure 3: Agent Internal View

Frequency) weighing scheme. Note that for the DF component of TF*IDF, values were computed within the information space of an agent. This was to follow the assumption that global information was not available to individuals and it is impossible to aggregate all documents in the network to get global DF values.

Provided TF*IDF representation, pair-wise similarity values were computed based on the Cosine coefficient that measures cosine of the angle between a pair of vectors [2]. For document clustering, we used the well-known K-means method based on cosine similarities [11].

Section 4.1 elaborates on a centroid-based method for query classification. Section 4.2 introduces a single-perceptron neural network (NN) algorithm for neighbor relevance prediction given query classification output. Section 4.3 discusses the formula for rewiring based on a *clustering exponent* α. For comparison, we also adopted a *Random Walk* model. The only difference was that in *Random Walk*, an agent simply ignored the neighbor selection step in Section 4.2 and forwarded a query to a random neighbor.

4.1 Centroid-based Query Classification

Given limited information each agent has, many widely appreciated classification methods, such as the Support Vector Machine (SVM), require a fair amount of training data and are therefore not applicable [23]. In this study, we used a simple centroid-based approach that produced competitive decentralized search results on a benchmark news collection [13].

Suppose an agent had k identified clusters/classes. Each class, $c \in [c_1, c_2..c_k]$, contained a set of documents $[d_1, d_2..d_n]$. Let $W_{d|i}$ denote the weight of the i^{th} term in document d. The weight of the i^{th} term in class centroid c was computed by:

$$W_{c|i} = \frac{\sum_{d=1}^{n_c} W_{d|i}}{n_c} \quad (1)$$

where n_c was the number of documents in class c. To classify a query, the query was first locally vectorized using the TF*IDF method and then compared to each class using the cosine similarity measure. The relevance of the classes to the query was sorted using the similarity scores.

4.2 Neural-Net for Neighbor Prediction

After query classification, the relevance (or similarity) of a query to each class was known. The topical relevance scores were then used to infer which neighbor was the best neighbor to contact if the current agent did not have relevant information. We assumed that the association between the classification output (a vector of topics' relevance scores) and the prediction (a vector of neighbors' relevance scores) is linear. A single perceptron neural network (NN) is suitable for the estimation of linear associations [23]. In this study, we implemented a feedforward perceptron NN with backprop and a sigmoid signal transfer function (please refer to [22] for details). To initialize learning, agents interact with their neighbors and learn about their topicality by using local documents as queries.

4.3 Peer Clustering and Network Rewiring

We introduced a clustering exponent α to rewire (reconnect peers through self-organization) a network and studied its impact on decentralized search. First, for each peer, some random peers were picked and added to its existing neighbors. Then, the current peer (i) queried all these neighbors (j) to determine their topical distance r_{ij} by sending them local documents as queries. Finally, the following connectivity probability function was used by the peer to decide who should remain as neighbors:

$$P_{ij} \propto r_{ij}^{-\alpha} \quad (2)$$

where α is the *clustering exponent* (or *homophily exponent*) and r_{ij} the pairwise topical distance. The finalized neighborhood size depended on the number of neighbors before rewiring. With a positive α value, the larger the topical distance, the less likely two peers will connect. Large α values lead to a highly clustered network while small values produce many topically remote connections or weak ties.

5. EXPERIMENTAL SETUP

We constructed a peer-to-peer network by using a large scholarly communication data collection and treating each unique scholar as a peer, who possessed a local collection of documents published by the scholar (author). The task involved finding a peer with relevant topic(s) in the network, given a query. Applications of this framework include, but are not limited to, distributed IR, P2P resource discovery, expert location in work settings, and reviewer finding in scholarly networks. However, we focused on the general decentralized search problem in large networked environments.

5.1 Data Collection

Data used in the experiments were from the TREC Genomics track 2004 benchmark collection, a Medline subset of about 4.5 million citations from 1994 to 2003. The data collection included metadata about publication titles, abstracts, and authors. We chose six scholars in the medical informatics domain and identified their direct co-authors (1^{st} degree) who published 10 to 80 articles in the TREC collection, resulting in a small network of 181 peers. Then the network was extended to the 2^{nd} degree (co-authors' co-authors) to total 5890 peers for experiments on a larger scale. Both networks had a diameter (the longest of all shortest pairwise paths) of 8 and roughly followed a power-law degree distribution with irregularities on the tail. For each peer, which represented a scholar/author, all articles (with titles and abstracts) authored or co-authored by the scholar were loaded as the local information collection.

5.2 Relevant Peers and Tasks

Relevant peers or information sources are considered few, if not rare, given a particular information need. To operationalize it, we defined a relevant peer as one of those who have the most similar information to a query. Specifically, we considered those scholars whose topical (cosine) similarity to a given query was ranked above the fifth percentile. Hence, for evaluation purposes, peers were sampled to estimate a threshold similarity score for each query, which was then used in experiments to judge whether a relevant peer had been found. We retrieved citations to articles published in the Journal of the American Medical Informatics Association (JAMIA) in the Genomics track collection and used all (498) articles with titles and abstracts as simulated queries.

5.3 Software and Hardware Setup

We developed a multi-agent system called RefNet, which takes advantage of the JADE [4] agent platform and the Weka machine learning framework [31]. RefNet has integrated the two major software packages (both in Java) to facilitate research experiments on decentralized search in networked environments.

Experiments were conducted on a Linux cluster of 9 nodes, each has Dual Intel Xeon e5405 (2.0 Ghz) Quad Core Processors (8 processors), 8 GB fully buffered system memory, and a Fedora 7 installation. The nodes were connected internally through a dedicated 1Gb network switch. The agents were equally distributed among the 72 processors, each of which loaded an agent container in Java, reserved 1GB memory, and communicated to each other. The Java Runtime Environment version for this study was 1.6.0_07.

5.4 Simulation Procedures

We ran experiments on the proposed RefNet model and a random-walk model and conducted comparative analyses. In both models, agents tried to forward a query to one another until one of the following conditions was met: 1) a relevant peer was found, or 2) the search path length reached its defined maximum. When concluded, the query would follow the search path in the reverse order back to the querying peer. Multiple runs were conducted in each parameter configuration. In each run, the 498 queries were submitted to the network one after another.

After experiments on initial co-authorship networks, we introduced the *clustering exponent* α to rewire the networks and studied its impact on decentralized search. Twenty random peers were added to each existing neighborhood, which was finalized based on the connectivity probability function defined in Section 4.3. It was further required that the final neighborhood size, for each peer, was in the range between 3 and 100.

5.5 Evaluation

The dependent variables of this study were effectiveness and efficiency of decentralized searches. We used completion rate of all tasks to measure retrieval effectiveness, $R_c = \frac{N_S}{N_T}$, where N_T is the total number of queries and N_S the number of them with a relevant peer found within given parameter limits.

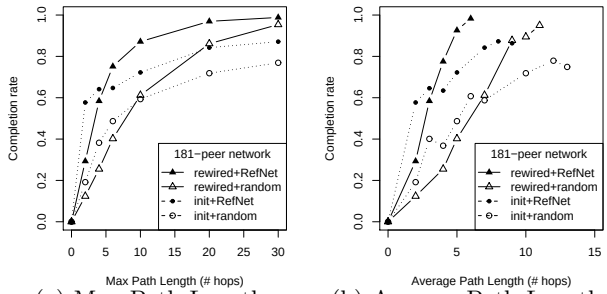
For efficiency, the maximum search path length L_{max} was controlled in each experiment and the actual path length of each task was measured. We computed average length of all searches in each experiment run, i.e., $\bar{L} = \frac{\sum_{i=1}^{N} L_i}{N_T}$, where L_i was the path length of the i_{th} query and N_T the total number of queries. With shorter path lengths, the entire distributed system is considered more efficient given fewer peers involved in computation.

For scalability, we ran experiments on different network sizes: 181 peers and 5890 peers. Effectiveness vs. efficiency patterns were compared. Various *clustering exponent* α values were controlled in experiments to examine its impact on the above variables. We further investigated the scaling of clustering impact in very large networks of up to 4 million peers based on synthetic data.

6. EXPERIMENTAL RESULTS

In this section, we present effectiveness and efficiency results on initial and rewired networks of 181 and 5890 peers, focus on the impact of clustering on decentralized search, and examine how the impact of network clustering scales.

6.1 181-Peer Network



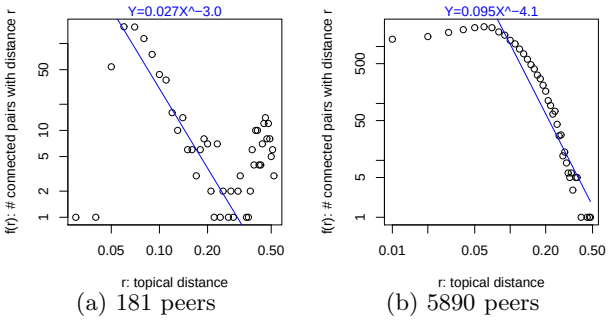
(a) Max Path Length (b) Average Path Length

Figure 4: Completion Rate (Y) vs. Path Length (X) on 181 Peers

Figure 4 shows experimental results on 181-peers networks. With the initial network (dotted lines), the *RefNet* model consistently outperformed random walks, especially within small path lengths. For instance, within two hops, RefNet already achieved a completion rate of more than 50% while random-walk was still at 20%. Increasing the path length helped both models but neither reached a completion rate higher than 90%, suggesting that there were particular characteristics of the initial network that disoriented some searches after a long path.

Clustering analysis, as plotted in Figure 5 (a) on log/log coordinates, showed that the association between connectivity frequency and topical distance has a power-law region (in the middle) with irregularities. We believe that *RefNet* searches were well guided by the network in most instances (when routed through peers with regular clustering-guided connections) but was lost in others (disoriented in regions where irregular connections dominated).

To demonstrate potential utility of network clustering, we rewired the network (throug self-organization) based on the connectivity probability function described in Section 4.3. Experimental results with clustering exponent $\alpha = 3.0$ are shown as solid lines in Figure 4, in which proper network clustering better guided RefNet search and further improved the results – a higher than 95% completion rate was already achieved at max search path length 20 (Figure 4 (a)) or average path length 5 (Figure 4 (b)).

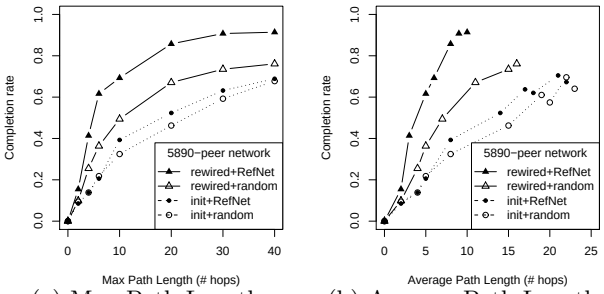


(a) 181 peers (b) 5890 peers

Figure 5: Initial Network Clustering: Connectivity (Y) vs. Topical Distance (X). Compare to Figure 1.

6.2 5890-Peer Network

On the initial 5890-peer network, experimental results indicated that the *RefNet* model had limited advantage over *random walk*, as shown by dotted lines in Figures 6 (a) and (b). Further analysis revealed that the network was insufficiently clustered. As shown in Figure 5 (b) on log/log coordinates, the correlation between connectivity and topical distance departed quite a bit from a power-law function (linear on log/log) with which efficient searches can be well-guided [15, 16, 24]. The curve suggests that there were too many topically remote connections that disoriented searches as peers were more likely to connect to topically irrelevant neighbors.



(a) Max Path Length (b) Average Path Length

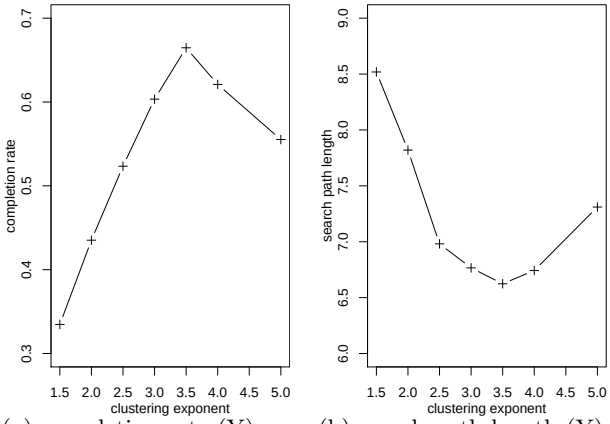
Figure 6: Completion Rate (Y) vs. Path Length (X) on 5890 Peers

Again, we used the method discussed in Section 4.3 to fine tune the 5890-peer network for a proper level of clustering. As shown by solid lines in Figure 6, given *clustering exponent* $\alpha = 4.0$, the *RefNet model* performed much better and achieved above 90% completion rate within a max path length of 40 (Figure 6 (a)) and with an average path length of about 10 (Figure 6 (b)).

6.3 Impact of Clustering

In the results above, we have demonstrated that some level of network clustering improved decentralized search of relevant peers or information sources. It is unclear yet how much clustering is enough or how much is too much. Setting max search path length at 10, experiments based on various clustering exponent α values on the 5890-peer network produced results shown in Figures 7 (a) and (b).

Given a constant max search path length at 10, Figure 7 (a) shows completion rate vs. clustering exponent α results,



(a) completion rate (Y) (b) search path length (Y)

Figure 7: Impact of Clustering Exponent α (X)

in which best completion rate was achieved at $\alpha \approx 3.5$, which also enabled optimal search path length in Figure 7 (b). Both smaller and larger α values resulted in less optimal searches. As discussed, smaller α values produced less visible topical segments and more remote connections that disoriented searches. Larger α values, on the other hand, led to an over-clustered and fragmented network without sufficient *weak ties* for searches to move fast.

This result, obtained in a decentralized information retrieval context, is consistent with findings from previous research on complex networks with simpler representations of the search (topical) space [15, 16, 24]. The *Clustering Paradox* suggests that when we use clustering-based techniques (e.g., topical segmentation and semantic overlay in P2P networks), some balance between *strong ties* and *weak ties* should be maintained.

Previous research also suggested that the optimal clustering exponent (the absolute value) is equal to the number of dimensions that describe topical distances among peers [15, 16]. We observed that the 181-peer network was optimal at $\alpha \approx 3.0$. With a larger number of peers and more diverse contents, the 5890-peer network seemed to require a little higher dimensionality to accurately depict all pairwise relationships, thus a slightly larger optimal clustering exponent $\alpha \approx 3.5$.

6.4 Scaling of Clustering Impact

One may argue that the impact of network clustering on decentralized search is small especially in the case of over-clustering – in Figure 7, for instance, there were roughly 10% loss in completion rate (effectiveness) and an increase of 1 in average search path length (efficiency) when α increased from 3.5 (optimum) to 5.0. Nonetheless, we will show in very large networks, the *Clustering Paradox* has a huge impact on search efficiency.

Relying on a 2-dimensional network model used in previous research [15, 16, 6], we ran decentralized search simulations on various network size scales $N \in [10^4, .., 4 \times 10^6]$ and with clustering exponent $\alpha \in [0, 4]$ (see [15] for detailed configurations). Results indicated that while optimum α approaches 2 with increased network size, there is a dramatic constrast between optimal clustering and overclustering in very large networks (see steeper curves in log-transformed Figure 8).

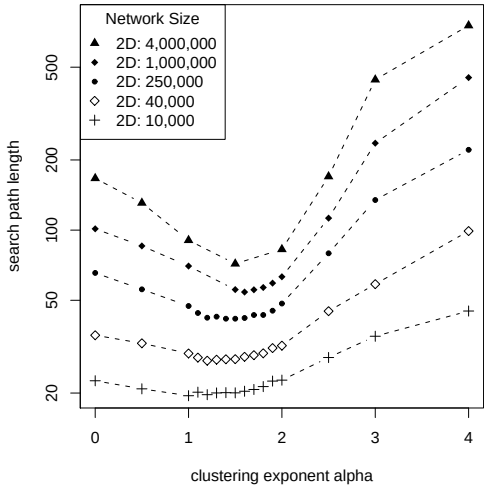


Figure 8: Scaling of Clustering Impact (100% completion rate). Note that search path length (Y) is log transformed.

On smaller scales (e.g., in the 10^4-peer network), as shown in Figure 8, optimization curves are much flatter. Overclustering in small networks only resulted in a moderate increase of search path length. However, in the network of four million peers, as shown in Figure 8, when α increased from 2 (nearly optimum) to 4, the average search path length increased from roughly 80 to more than 700 – a huge loss in search efficiency. Seen in this light, methods achieving good results on small or medium network sizes will not necessarily function well on large scales. Little performance disadvantage in small networks might become too big to ignore in large networks. Scrutiny of the *Clustering Paradox* for network optimization is crucial for scalability of decentralized search.

7. CONCLUSION

In this paper, we presented a multi-agent framework for information retrieval in distributed networked environments and focused on the impact of network clustering on decentralized search. Particularly, we studied search optimization in the face of the *Clustering Paradox*, in which either too little or too much clustering leads to degraded findability of relevant information sources. Experiments showed that the similarity based *RefNet* model outperformed random walks on the initial 181-peer network and did not show much advantage on the initial 5890-peer network, which was shown to have too many topically remote connections or *weak ties* that disoriented searches.

By introducing a *clustering exponent* α to guide network rewiring, we studied the impact of clustering and found that a balanced level of network clustering produced optimal results. Particularly, in the network of 5890 scholars, relevant peers were best findable at $\alpha \approx 3.5$. Smaller α values resulted in less visible topical segments and many remote connections that disoriented searches. Larger α values, on the other hand, led to an over-clustered and fragmented network with rich *strong ties* but scant *weak ties* for searches to move fast.

Further experiments on various larger networks of up to 4 million peers demonstrated that clustering optimization is crucial for decentralized search. Although overclustering only moderately degraded search performance on small networks, it led to dramatic loss in search efficiency for large networks. So did weak clustering. Search methods that work well on small scales might function badly in large networks, in which little performance disadvantage in small networks might become too big to ignore. As many research rely on clustering for decentralized search (e.g., in semantic overlay networks for P2P), scrutiny of the *Clustering Paradox* is crucial for scalability of existing methods.

Acknowledgments

We appreciate valuable discussions with Gary Marchionini, Munindar P. Singh, Diane Kelly, Jeffrey Pomerantz, and Simon Spero, and constructive comments from LSDS-IR'09 reviewers. We thank the NC Translational and Clinical Sciences (TraCS) Institute for support.

8. REFERENCES

[1] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149–11152, 2000.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman Publishing, 2004.

[3] M. Bawa, G. S. Manku, and P. Raghavan. Sets: search enhanced by topic segmentation. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 306–313, New York, NY, USA, 2003. ACM.

[4] F. L. Bellifemine, G. Caire, and D. Greenwood. *Developing Multi-Agent Systems with JADE (Wiley Series in Agent Technology)*. John Wiley & Sons, 2007.

[5] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving collection selection with overlap awareness in p2p search engines. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–74, New York, NY, USA, 2005. ACM.

[6] M. Boguñá, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nature Physics*, 5(1):74 –80, 2009.

[7] A. Crespo and H. Garcia-Molina. Semantic overlay networks for p2p systems. In *Agents and Peer-to-Peer Computing*, pages 1–13, 2005.

[8] C. Doulkeridis, K. Norvag, and M. Vazirgiannis. Peer-to-peer similarity search over widely distributed document collections. In *LSDS-IR '08: Proceeding of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval*, pages 35–42, New York, NY, USA, 2008. ACM.

[9] G. Fischer and A. Nurzenski. Towards scatter/gather browsing in a hierarchical peer-to-peer network. In *P2PIR '05: Proceedings of the 2005 ACM workshop on Information retrieval in peer-to-peer networks*, pages 25–32, New York, NY, USA, 2005. ACM.

[10] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, May 1973.

[11] J. Han, M. Kamber, and A. L. H. Tung. *Spatial Clustering methods in data mining: a survey*. CRC, New York, 2001.

[12] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pages 76–84, New York, NY, USA, 1996. ACM Press.

[13] W. Ke, J. Mostafa, and Y. Fu. Collaborative classifier agents: studying the impact of learning in distributed document classification. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 428–437, New York, NY, USA, 2007. ACM.

[14] W. Ke, C. R. Sugimoto, and J. Mostafa. Dynamicity vs. effectiveness: Studying online clustering for scatter/gather. In *SIGIR '09: Proceedings of the 32th annual international ACM SIGIR conference on research and development in information retrieval*, Boston, MA, 2009. ACM Press.

[15] J. M. Kleinberg. Navigation in a small world. *Nature*, 406(6798), August 2000.

[16] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[17] J. Lu and J. Callan. User modeling for full-text federated search in peer-to-peer networks. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 332–339, New York, NY, USA, 2006. ACM.

[18] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim. A survey and comparison of peer-to-peer overlay network schemes. *IEEE Communications Surveys and Tutorials*, 7:72–93, 2005.

[19] T. Luu, F. Klemm, I. Podnar, M. Rajman, and K. Aberer. Alvis peers: a scalable full-text peer-to-peer retrieval engine. In *P2PIR '06: Proceedings of the international workshop on Information retrieval in peer-to-peer networks*, pages 41–48, New York, NY, USA, 2006. ACM.

[20] P. Raftopoulou and E. G. Petrakis. A measure for cluster cohesion in semantic overlay networks. In *LSDS-IR '08: Proceeding of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval*, pages 59–66, New York, NY, USA, 2008. ACM.

[21] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A scalable content-addressable network. In *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 161–172, New York, NY, USA, 2001. ACM.

[22] R. D. Reed and R. J. Marks. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press, Cambridge, MA, USA, 1998.

[23] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[24] O. Simsek and D. Jensen. Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, 105(35):12758–12762, 2008.

[25] M. P. Singh, B. Yu, and M. Venkatraman. Community-based service location. *Communications of the ACM*, 44(4):49–54, 2001.

[26] G. Skobeltsyn, T. Luu, I. P. Zarko, M. Rajman, and K. Aberer. Web text retrieval with a p2p query-driven index. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 679–686, New York, NY, USA, 2007. ACM.

[27] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 149–160, New York, NY, USA, 2001. ACM.

[28] C. Tang, Z. Xu, and S. Dwarkadas. Peer-to-peer information retrieval using self-organizing semantic overlay networks. In *SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 175–186, New York, NY, USA, 2003. ACM.

[29] C. J. van Rijsbergen and K. Sparck-Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, 1973.

[30] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and Search in Social Networks. *Science*, 296(5571):1302–1305, 2002.

[31] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[32] B. Yu and M. P. Singh. Searching social networks. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 65–72, New York, NY, USA, 2003. ACM.

[33] I. P. Zarko and F. Silvestri. The CIKM 2006 workshop on information retrieval in peer-to-peer networks. *SIGIR Forum*, 41(1):101–103, 2007.

[34] D. Zeinalipour-Yazti, V. Kalogeraki, and D. Gunopulos. Information retrieval techniques for peer-to-peer networks. *Computing in Science and Engineering*, 6(4):20–26, 2004.

[35] H. Zhang and V. Lesser. A reinforcement learning based distributed search algorithm for hierarchical peer-to-peer information retrieval systems. In *AAMAS '07: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–8, New York, NY, USA, 2007. ACM.

[36] J. Zhang and M. S. Ackerman. Searching for expertise in social networks: a simulation of potential strategies. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 71–80, NY, USA, 2005. ACM.