Performance Bounds of Decentralized Search in Expert Networks for Query Answering

LIANG MA and MUDHAKAR SRIVATSA, IBM T. J. Watson Research
DERYA CANSEVER, Army Research Laboratory
XIFENG YAN, University of California, Santa Barbara
SUE KASE and MICHELLE VANNI, Army Research Laboratory

Expert networks are formed by a group of expert-professionals with different specialties to collaboratively resolve specific queries posted to the network. In such networks, when a query reaches an expert who does not have sufficient expertise, this query needs to be routed to other experts for further processing until it is completely solved; therefore, query answering efficiency is sensitive to the underlying query routing mechanism being used. Among all possible query routing mechanisms, decentralized search, operating purely on each expert's local information without any knowledge of network global structure, represents the most basic and scalable routing mechanism, which is applicable to any network scenarios even in dynamic networks. However, there is still a lack of fundamental understanding of the efficiency of decentralized search in expert networks. In this regard, we investigate decentralized search by quantifying its performance under a variety of network settings. Our key findings reveal the existence of network conditions, under which decentralized search can achieve significantly short query routing paths (i.e., between $O(\log n)$ and $O(\log^2 n)$ hops, n: total number of experts in the network). Based on such theoretical foundation, we further study how the unique properties of decentralized search in expert networks are related to the anecdotal small-world phenomenon. In addition, we demonstrate that decentralized search is robust against estimation errors introduced by misinterpreting the required expertise levels. The developed performance bounds, confirmed by real datasets, are able to assist in predicting network performance and designing complex expert networks.

CCS Concepts: • **Information systems** → **Expert search**; • **Human-centered computing** → **Social networks**; • **Networks** → *Network performance analysis*;

Additional Key Words and Phrases: Expert networks, query answering, decentralized search, performance bounds, theory

ACM Reference format:
Liang Ma, Mudhakar Srivatsa, Derya Cansever, Xifeng Yan, Sue Kase, and Michelle Vanni. 2019. Performance Bounds of Decentralized Search in Expert Networks for Query Answering. *ACM Trans. Knowl. Discov. Data* 13, 2, Article 18 (March 2019), 23 pages.
https://doi.org/10.1145/3300230

This work was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA).

Authors' addresses: L. Ma and M. Srivatsa, IBM T. J. Watson Research, 1101 Kitchawan Rd, Yorktown, NY, 10598; emails: {maliang, msrivats}@us.ibm.com; D. Cansever, Army Research Laboratory, 2800 Powder Mill Rd, Adelphi, MD, 20783; email: derya.h.cansever.civ@mail.mil; X. Yan, University of California, Santa Barbara, Santa Barbara, CA, 93106; email: xyan@cs.ucsb.edu; S. Kase and M. Vanni, Army Research Laboratory, 2800 Powder Mill Rd, Adelphi, MD, 20783; emails: {sue.e.kase.civ, michelle.t.vanni.civ}@mail.mil.

1556-4681/2019/03-ART18 $15.00
https://doi.org/10.1145/3300230

1 INTRODUCTION

Expert networks are composed of a group of expert-professionals, who cooperate with each other to solve specific queries (e.g., reported by clients) using their professional knowledge in a variety of related subjects. Such expert networks are abundant in real life especially in commercial organizations, where networks with specialized experts are maintained to provide consulting/troubleshooting services. The collaboration among experts is knowledge-driven, manifesting in the process of expert searching: Upon receiving a query by an expert, she first attempts to solve the problem specified in the query; if she fails, then this query is routed to another expert for further processing. This process continues until the query is resolved. One canonical example of expert networks is the enterprise call center. Specifically, in a call center, if a query ticket cannot be solved by the first responding agent, then a series of processing/forwarding attempts are triggered until qualified agents are found. The fundamental goal regarding expert networks is to route each query to experts with sufficient expertise in a timely and accurate manner. This is a challenging issue as efficient query routing mechanisms depend on the professional knowledge of each individual expert as well as the social knowledge of other experts' specialties possessed at each expert. When the expert profiles (e.g., expertise) are not properly maintained in expert networks, finding the most knowledgeable experts with high probability while minimizing the number of routing steps is explored in [2, 29]. On the other hand, even if each expert's profile is accurately exposed to all of her contacts, this routing issue remains challenging. In particular, under the assumption that each expert has connections to only a limited number of other experts, a series of routing rules are proposed in [30, 40] for improving the resolution efficiency in specific tasks (e.g., IT services). Under the same assumption, generative models [25] are developed for making global expert recommendations by estimating all possible routes to potential resolvers. The efficacy of these mechanisms, however, highly relies on the broad knowledge of network global structure; in addition, these mechanisms, requiring non-negligible training periods, are generally complicated and sensitive to operational scenarios, thus not applicable to large-scale or dynamic networks (e.g., with experts joining/leaving the network). All these limitations, therefore, motivate us to consider if there exist simple yet efficient query routing solutions that function with only basic network information and are robust against network variations.

Among all possible query routing mechanisms, *decentralized search*, operating purely on each expert's local information, represents the most basic and adaptive routing mechanism in expert networks. Specifically, under decentralized search, before reaching experts with sufficient expertise for a given query, each intermediate expert forwards this query to one of her contacts with the highest problem solving abilities, which therefore forms a pure local-information-based forwarding rule. Since no historical training data or network global structure is required for routing decision making, decentralized search can be broadly adopted by any network scenarios for query routing. However, decentralized search, greedy in nature at each routing step, is generally ignored in the research community mainly for the following reason: When an expert forwards a query to one of her contacts via decentralized search, she has no clue whether this decision would successfully lead to a short path through the entire network, and thus the efficiency of decentralized search is uncertain. We note, however, without fundamental understanding of such simple decentralized search, we can never justify the value/necessity of designing other complicated query routing mechanisms for expert networks. Therefore, in this article, we consider this unsolved fundamental problem: What is the efficiency of decentralized search in expert networks? We study this problem by quantifying the performance of decentralized search under various network settings so as to understand under what conditions decentralized search can achieve efficiency/inefficiency in query routing.

In this article, the basic approach we employ to study the performance of decentralized search is to establish its performance bounds in generic expert network models. Such models should capture two main connection properties among experts as follows: (i) experts are rich in connections to peer experts with similar expertise; (ii) each expert also tends to connect to a few experts with fairly dissimilar expertise. Integrating these two properties that characterize expert social connections, we propose two expert network models. In these models, local expert connections (experts with similar expertise) enable the formation of the basic network structure, on top of which long-range expert connections (experts with fairly dissimilar expertise) determine to what extent the expert inter-connections do not respect such basic network structure. We prove that the natural superposition of these two properties in expert networks can lead to high efficiency of decentralized search without any centralized guidance under a range of network settings. Accordingly, if an expert network is verified to satisfy such conditions that guarantee efficient decentralized search, then there is no need to design complicated query routing mechanisms as the light-weighted decentralized search will suffice. Furthermore, by a case study of real datasets, we demonstrate how commercial expert networks may take proactive actions to train their constituent experts, which equivalently approaches the efficient query routing conditions discussed in this article.

To gain more insights into decentralized search, we then consider how its efficiency relates to the intriguing and pervasive *small-world phenomenon* [23, 26, 32], a principle stating that any two individuals in the network are connected by a short chain of intermediate acquaintances. Note that as opposed to observing small-world phenomenon in expert networks, we aim to study, using the above theoretical results, the relationship between the existence of short routing chains and the efficacy of decentralized search in finding such short chains, thus providing better understanding of the uniqueness of small-world navigation via decentralized search in expert networks.

1.1 Further Discussions on Related Work

Expert networks essentially are graphs where nodes are associated with descriptions (i.e., expertise profiles); therefore, there are many works on expert networks from the graph embedding perspective, with the goal to learn node vector representations that capture the network properties of interest. Specifically, node2vec [13] designs flexible node sampling methodologies to allow feature vectors to exhibit different properties. Subgraph2vec [27] extends these schemes to learn vector representations of subgraphs. When network nodes are associated with auxiliary information (e.g., node expertise profiles), the work in [15], [36], and [17] further address the case where nodes are associated with labels. [28] studies graph embedding when node/edge attributes are continuous. The author in [7] investigates phrase ambiguity resolution via leveraging hyperlinks. However, all these works operate under information or network constraints. On the other hand, the author in [35], [38], and [33] explore embedding strategies in the context of knowledge graphs, where the main goal is to maintain the entity relationships specified by semantic edges. In contrast, in expert networks, only nodes are associated with semantic information. In this regard, EP [12] and GraphSAGE [16] learn embeddings for structured graph data. However, the textual similarities are only captured by linear combinations. Planetoid [37] computes node embeddings under semi-supervised settings; metapath2vec [10] learns embeddings for heterogeneous networks (node can be author or paper); and Graph-Neural-Network-based embeddings are explored in [18] and [24]. However, these papers only provide solutions to capture the structural properties and/or nodes' descriptions in expert networks. In other words, the query routing in these networks are not explored. By contrast, the objective of this article is not only to study the query routing in expert networks, but also to quantify its performance from the analytical perspective.

Regarding query routing in expertise networks, most existing works [2, 29, 30, 40] seek to develop/improve query routing mechanisms, where different levels of network global knowledge are required. With network historical data, the author in [3] develops a Markov Decision Process (MDP) model to optimize routing policies. However, the correlation between successful query answering probability and each expert's expertise level is ignored in the proposed model. Employing game theory, the author in [22] proposes query incentive networks to understand agent collaborations and interactions in on-line communities. In addition, routing efficiency improvement is investigated in [39] when additional expert contacts are carefully chosen. Our work belongs to a different but closely related line of work that focuses on the fundamental understanding of the most basic query routing mechanism in expert networks. Our work shares similar goals with [31] in that [31] tries to build models to capture routing behaviors in expert networks, particularly in modeling human factors that influence routing tendencies. However, the author in [31] does not show how the efficiency of such routing behaviors are affected by network and social properties, and no performance guarantee is investigated. By contrast, we not only present deep insights into decentralized search (which is able to capture routing behaviors in some real networks; see the case study in Section 6) in expert networks, but also show how its efficacy is related to network's structural and social properties.

For the underlying influence of network properties on the routing efficiency, the authors in [5] and [4] show that every pair of nodes are joined by a path of length $O(\log n)$ (n: total number of nodes in the network) in a randomly generated graph. The existence of such short paths is maintained even when the network demonstrates certain structural properties [34]. Further, more complicated statistical models are considered in [11] for node inter-connections (e.g., Poisson distributions). When the number of links incident to nodes follows the power-law distribution, the work in [1] explores how such distribution may affect routing preference. In addition, with special network properties, networks may exhibit small-world phenomenon [23, 26, 32]. The intriguing characteristic of small-world phenomenon has stimulated numerous compelling research results [9, 14, 19–21, 34], among which [20] is the first work showing that there exists one and only one network setting that enables efficient searching algorithms. In this article, we do not seek to observe small-world phenomenon in another type of networks (i.e., expert networks); by contrast, we aim to understand the small-world phenomenon with characteristics that uniquely exist in expert networks using our fundamental theoretical results on decentralized search. To this end, we prove that the anycast nature of query routing (i.e., the number of qualified experts may be more than one) can lead to decentralized search being highly efficient under a wide range of network settings, which is completely different from prior works on the small-world phenomenon.

1.2 Summary of Contributions

We study, for the first time, decentralized search in expert networks from the perspective of fundamental performance quantifications. Our contributions are seven-fold:

(1) We build mathematical models to formulate abundant expert connections to similar experts and a few connections to dissimilar experts, in terms of their expertise differences.
(2) To capture the two properties in (1), we propose two expert network models: (i) unified model, where all experts have the same overall problem solving abilities, but specialize in different areas, and (ii) diversified model, where experts may have different per-area expertise or different overall problem solving abilities. In the diversified model, the per-expert total problem solving ability exhibits a Gaussian-like distribution as the number of solvable subjects in the network increases.
(3) We prove that decentralized search is highly efficient under a wide range of network settings for both unified and diversified models; the corresponding average routing path length is between $O(\log n)$ and $O(\log^2 n)$ (n: total number of experts).

(4) We further establish conditions for the case when decentralized search is ineffective, and develop the corresponding lower bounds to quantify its performance.
(5) We discuss how above theoretical results are related to the special characteristics of small-world phenomenon in expert networks. We reveal that the existence of small-world phenomenon (under wide conditions) directly leads to efficient decentralized search in expert networks. However, in point-to-point (unicast) networks, one and only one of these conditions enables efficient local-information-based search.
(6) We demonstrate that decentralized search is robust in the case where experts experience interpretation errors regarding the expertise requirement in the received queries.
(7) We show that the above theoretical bounds can also approximate the routing performance in real datasets, even though the network structures in these real datasets may not rigorously respect the proposed network models; therefore, these theoretical results can provide guidance in planning practical complex expert networks.

Note that in this article, the focus is the fundamental understanding of query routing behaviors under the assumption that experts' expertise and social connections are fixed. We acknowledge that both professional and social knowledge of experts may improve over time, thus benefiting the query routing efficiency. In such case where improved routing information is available, the above results remain valid as long as the network parameters in the newly formed expert network (e.g., experts' improved expertise and richer interconnections) are retrieved and updated (see the case study in Section 6 where the routing efficiency is compared in networks with expertise and connection evolvement).

The remainder of the article is organized as follows. Section 2 formulates the problem. Two models for expert networks are proposed in Section 3. Main results of decentralized search in expert networks are presented and analyzed in Section 4, where the corresponding proofs are shown in Section 5. Experiments are conducted under both synthetic networks and real datasets in Section 6. Finally, Section 7 concludes the article.

2 PROBLEM FORMULATION

In this section, we propose mathematical models to capture expert inter-connections, and then formally present decentralized search and state our research objective.

2.1 Expert Inter-Connections

We assume that in an expert network with n experts, experts can collectively solve problems in up to m different areas. To model such an expert network, we use directional edges to represent expert inter-connections. In particular, expert u can route a query to expert w if and only if there exists a directional edge from u to w, denoted by $\overrightarrow{uw}$, where w is called a *contact* of u. For all experts, we assume that their expertise in different areas are quantifiable to non-negative integers, and thereby each expert is associated with an expertise vector defined as follows: The *expertise vector* of expert u, denoted by $\mathbf{e}^{(u)}$, is an $m \times 1$ column vector with the value in entry i (i.e., $e_i^{(u)}$) indicating u's skill in area i (larger value corresponds to superior skill); $e_j^{(u)} = 0$ if u does not have any skill in area j. We call $||\mathbf{e}^{(u)}||_1 := \sum_i |e_i^{(u)}|$ the *total ability* of expert u. Using this concept, we can compare the expertise levels in different areas for one individual expert or in the same area across multiple experts. Furthermore, we define the *expertise distance* from expert u to expert w as $d(u \to w) := \sum_i \max(e_i^{(w)} - e_i^{(u)}, 0)$. Intuitively, expertise distance characterizes the superior skills of one expert against another, and it implies that generally $d(u \to w) \neq d(w \to u)$. Moreover, when $d(u \to w) = 0$, it does not mean u and w have similar expertise; on the contrary, it only suggests that $\mathbf{e}^{(u)} \geq \mathbf{e}^{(w)}$, i.e., u is superior than w in all expertise areas; see later discussions on how expertise distance is used in constructing expert inter-connections. With all these concepts, we are ready to model homophily and heterophily of expert inter-connections.

Homophily refers to the tendency that each expert is rich in connections to peer experts with similar expertise. To characterize such expertise similarity, both inferior and superior skills should be considered when comparing two experts, i.e., expertise difference in all areas between two experts must be within a threshold. Therefore, a natural way to model homophily is as follows: For a universal constant integer $\delta \geq 1$, called *similarity degree*, each expert (denoted by u) connects to all experts in set $R := \{w \in V \setminus \{u\} : ||\mathbf{e}^{(w)} - \mathbf{e}^{(u)}||_1 \leq \delta\}$, where V is the set of experts in the entire network and "\" is set minus. All experts in set R are called *local contact*s of expert u. When experts are connected by the homophily rule (adding local contacts for each expert), an expert network basis, called *network substrate*, is formed. The network substrate is determined purely by the constant parameter δ, and thus the network substrate does not exhibit any randomness (assuming that the expertise vectors of all experts are fixed).

Heterophily refers to the phenomenon that each expert has a few connections to experts with fairly dissimilar expertise. These dissimilar experts are called *long-range contact*s, which are crucial in determining the network diameter (average length of the shortest path connecting each pair of nodes). However, unlike homophily, there exists randomness in observing which two dissimilar experts are connected. Therefore, we use a statistical model to capture heterophily, which consists of the following two steps: (i) for expert u, compute the set of candidates C_u that can be long-range contacts of u by $C_u := \{w \in V \setminus \{u\} : \mathbf{e}^{(w)} \not\leq \mathbf{e}^{(u)}\}$; (ii) for a universal constant integer $k \geq 1$ and constant $r \geq 0$, expert u has k out-going edges to connect to long-range contacts with independent probabilities. Specifically, for the k out-going edges from u, each edge terminates at expert w ($w \in C_u$) with probability, denoted by $\Pr(u \rightarrow w)$, proportional to $[d(u \rightarrow w)]^{-r}$. In other words, $\Pr(u \rightarrow w) = [d(u \rightarrow w)]^{-r} / \sum_{v \in C_u} [d(u \rightarrow v)]^{-r}$, called the *inverse rth power distribution.*

In practical expert networks, the goal of having long-range contacts is for efficient query routing; therefore, one expert's long-range contact must exhibit superior expertise in certain areas, as otherwise such long-range contact is useless for routing. Hence, this heterophily model captures that expert u connecting to a long-range expert w if and only if w has superior skills in certain areas compared to u; therefore, expertise distance $d(u \rightarrow w)$ can be used to capture such expertise dissimilarity. Moreover, when $d(u \rightarrow w) = 0$, w is not a contact of u for the query routing task; nevertheless, since all connection edges are directional, u might be a contact of w for query routing as it is possible that $d(w \rightarrow u) > 0$. Note that the long-range contact construction rule indicates that the ith and jth edges from u may terminate at the same expert (may even be one of u's local contacts); therefore, k is the maximum number of long-range contacts for each expert. Moreover, r serves as a structural parameter that controls the scale of long-range contacts for each individual expert. In particular, when $r = 0$, all experts in the candidate set C_u are equally likely to be long-range contacts of u, which corresponds to a purely random case; when r increases, long-range contacts of an expert tend to only exist within her vicinity (measured by the expertise distance); when r approaches $+\infty$, all long-range contacts disappear, i.e., there is no heterophily in the expert network. In this article, we study how these parameters relate to the efficiency of decentralized search. All notations used in this article are summarized in Table 1.

Illustrative Example: Figure 1(a) shows a sample network with six experts, where the expertise distance between two experts is also the difference of their IDs. In Figure 1(a), each expert has the most similar experts as local contacts. The probability of Expert 1 having long-range contacts with other experts is sketched in Figure 1(b), which shows that when r increases, the impact of expertise distances becomes salient, thus causing decreased connection probability.

Table 1. Notations

Symbol	Meaning
V, n	Set/number of experts ($n = \|V\|$)
r	Long-range contact follows an inverse rth power distribution
C_u	Set of candidates who can become long-range contacts of u
k	Maximum number of long-range contacts for each expert
h	Number of rows in the unified model
m	Number of elements in an expertise vector
$\mathbf{e}^{(u)}$	Expertise vector ($m \times 1$ column vector) of expert u
(i, τ)	Query in problem area i with difficulty level being τ (query (i, τ) is solvable by an expert u with $e_i^{(u)} \geq \tau$)
λ	Maximum expertise level in each area for any expert in the diversified model
δ	Similarity degree
c	Scaling factor in the standard deviation of query interpretation: $\sigma = c(\tau - e_i^{(u)})$ for query (i, τ) at expert u
$d(u \rightarrow w)$	$d(u \rightarrow w) = \sum_i \max(e_i^{(w)} - e_i^{(u)}, 0)$ expertise distance from u to w

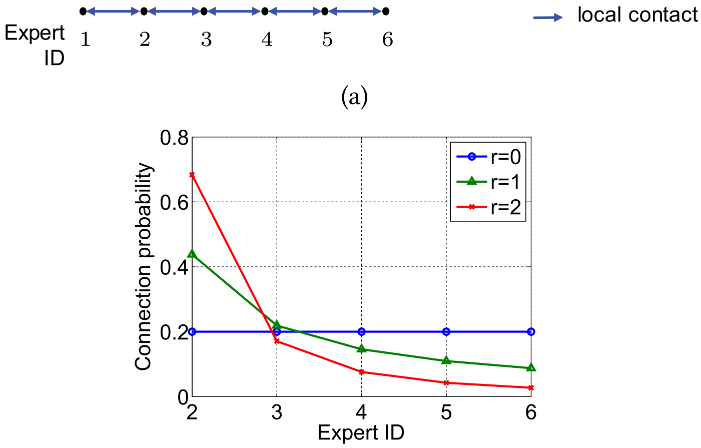


Fig. 1. Illustrative example: (a) sample network (expertise distance from Expert i to Expert $i + j$ is j); (b) probability of Expert 1 having other experts in (a) as long-range contacts under different values of r.

2.2 Decentralized Search

We now formally present decentralized search for query routing. For the queries posted to the expert network, they are generally first categorized into problem areas, and then these queries are routed to one or more experts with sufficient expertise to resolve. In this article, to simplify the problem formulation, we consider the case where each received query belongs to one and only one problem area. If a query contains problems in p ($p > 1$) areas, then this query will be treated as p separate queries. In this regard, we model each query as a 2-tuple (i, τ), where i is the problem area to which this query belongs and τ ($\tau > 0$) indicates the corresponding difficulty level, i.e., query (i, τ) is solvable by experts with expertise level in area i being at least τ. We assume that there is no ambiguity in determining the problem areas of queries, and there exist

ALGORITHM 1: Decentralized Search

```
input: Expert network, query $(i, \tau)$, first query holder $u$
output: Routing path $\mathcal{P}$ for resolving query $(i, \tau)$
$\mathcal{P} \leftarrow u$;                                  // "←":assignment operation
while $e_i^{(u)} - \tau < 0$ do
    $u = \arg\max_{w \in \mathcal{N}(u)} \left( \min(e_i^{(w)} - \tau, 0) \right)$ ; // $\mathcal{N}(u)$: set of all (local and long-range) contacts of $u$
    $\mathcal{P} \leftarrow \mathcal{P} + u$ ;                        // append $u$ to $\mathcal{P}$
end
```

qualified experts in the network to solve each query, i.e., for any query (i, τ), $\exists$ expert w with expertise in area i and $e_i^{(w)} \geq \tau$. In this article, the most crucial assumption is that for each query holder, besides knowing the expertise vector and (local and long-range) contacts of herself, she also knows the expertise vectors of all her (local and long-range) contacts; however, she does not have knowledge of expertise vectors or contacts of other experts, i.e., no experts have the global view[1] of the network. Under these assumptions, decentralized search is detailed in Algorithm 1. In Algorithm 1, for a given query, if the current query holder u cannot solve this query, then line 5 searches for the best expert from all u's contacts (with ties broken arbitrarily) as the next routing step. This process continues until a qualified expert is found.

Remark: In decentralized search, if a query holder's contacts cannot solve the received query, then this query holder has no knowledge of where the qualified experts are. Therefore, one may concern that the condition in line 4 may never be satisfied for some queries, thus resulting in endless loops. We will show in Section 3 that the structural properties of the two representative network models abstracted from real networks ensure that at least one expert satisfying the condition in line 4 can be found by Algorithm 1 (although the resulting routing path may be long).

2.3 Objective

Suppose that the problem area and the difficulty level in each query are independently and uniformly distributed (subject to the maximum problem solving ability in the network) and the first query holder is also randomly chosen. Our goal is to understand decentralized search in expert networks by computing its upper/lower bound of the *average routing path length* (measured by the number of hops) under different network structures and expert inter-connections.

3 EXPERT NETWORK MODELS

Based on expert inter-connection models in Section 2, we now present two expert network models, i.e., unified and diversified models, which differ by the distribution of expert total abilities. The significance of these models is that each of them captures unique features in real expert networks.

3.1 Unified Model

The first network model is called the *unified model*, where all experts have the same total abilities, i.e., the sum of all entries in the expertise vector of any expert is a constant, though expertise in different areas may vary. The unified model captures the main features in real expert networks where experts have (almost) the same problem solving abilities (e.g., within the same organization or with similar payment), yet also have their own specialties (e.g., they are hired according to

[1] If the global picture of the network is fully known to each expert, then simple breadth-first search will suffice to find the shortest query routing path, which is not of interest to this article.

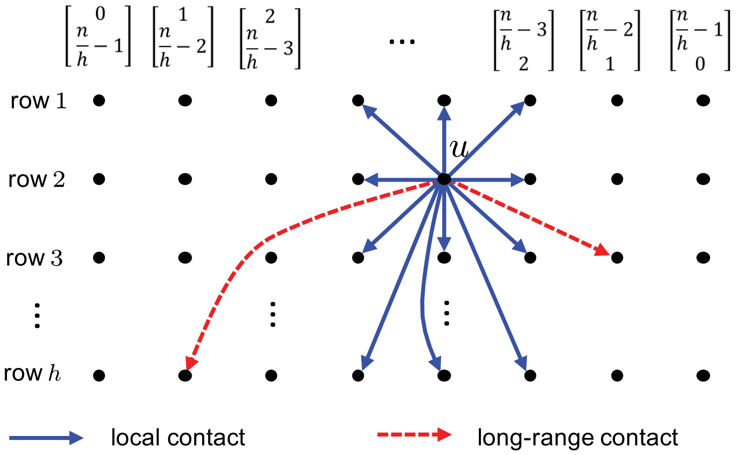

$\begin{bmatrix}0\\ \frac{n}{h}-1\end{bmatrix}$
$\begin{bmatrix}1\\ \frac{n}{h}-2\end{bmatrix}$
$\begin{bmatrix}2\\ \frac{n}{h}-3\end{bmatrix}$
$\cdots$
$\begin{bmatrix}\frac{n}{h}-3\\ 2\end{bmatrix}$
$\begin{bmatrix}\frac{n}{h}-2\\ 1\end{bmatrix}$
$\begin{bmatrix}\frac{n}{h}-1\\ 0\end{bmatrix}$
row 1
row 2
row 3
row h
u
local contact
long-range contact

Fig. 2. Unified model ($\delta = 2$, $k = 2$, $[\cdot]$: expertise vector).

company's multi-development goals). Suppose the expert network can solve queries in up to two specific areas (i.e., $m = 2$), then the unified model is structured as follows:

(1) n experts are distributed in an $h \times \frac{n}{h}$ grid (h rows and n/h columns as shown in Figure 2, assuming n/h is an integer);
(2) experts in the same column have the same expertise vector;
(3) in each row, expertise in the first area increases from 0 to $\frac{n}{h} - 1$ by 1 at each expert in the direction from left to right, while the expertise in the second area increases from 0 to $\frac{n}{h} - 1$ by 1 at each expert in the opposite direction;
(4) w.r.t. each expert in a row, she has the most similar experts (i.e., similarity degree $\delta = 2$) as local contacts. Thus, each expert with expertise vector $[0, (n/h) - 1]^T$ or $[(n/h) - 1, 0]^T$ has $2h - 1$ local contacts; all other experts each has $3h - 1$ local contacts (see Figure 2).

Then, based on the above network substrate, long-range contacts are constructed following the inverse rth power distribution (see Section 2) for each expert; see Figure 2.

Discussions: In practical networks where all experts have similar expertise levels, the network structure may not strictly follow such unified model. However, we can simplify the given network and use the minimum or the maximum number of experts in each column as the value of h (h rows in the unified model) so as to estimate the performance bound of the decentralized search. Moreover, this unified model can be extended to cases where $m > 2$ (i.e., the network can solve problems in more than two areas). Such extension will be discussed after presenting the diversified model as one subgraph in the diversified model is required for extending the unified model to the case where $m > 2$. Finally, we point out that the significance of the unified model presented in this article is that it serves as the fundamental building block for complex models with experts all having the same total abilities.

3.2 Diversified Model

The second network model is called the *diversified model*, in which both the total abilities and specialties may vary for different experts. One advantage of this model is that it naturally captures the Gaussian-like distribution of expert total abilities in real expert networks. Suppose up to m ($m \geq 1$) specific areas can be solved in the expert network. Let λ denote the maximum expertise level in

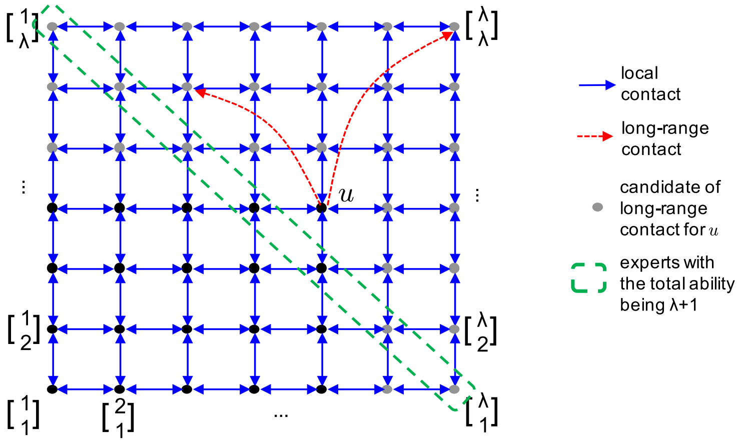

$\begin{bmatrix}1\\\lambda\end{bmatrix}$
$\begin{bmatrix}\lambda\\\lambda\end{bmatrix}$
⋮
u
⋮
$\begin{bmatrix}1\\2\end{bmatrix}$
$\begin{bmatrix}\lambda\\2\end{bmatrix}$
$\begin{bmatrix}1\\1\end{bmatrix}$
$\begin{bmatrix}2\\1\end{bmatrix}$
...
$\begin{bmatrix}\lambda\\1\end{bmatrix}$
local contact
long-range contact
candidate of long-range contact for u
experts with the total ability being λ+1

Fig. 3. Diversified model ($\delta = 1, m = 2, k = 2$, $[\cdot]$: expertise vector).

each area (i.e., the maximum value for every entry in any expertise vector), then the diversified model is structured as follows:

(1) in the network with n experts, let $\lambda := \sqrt[m]{n}$. Suppose for any expertise vector $\mathbf{e}$, $\forall i$, e_i is an integer between 1 and λ (i.e., $[1, 1, \ldots, 1]^T \preceq \mathbf{e} \preceq [\lambda, \lambda, \ldots, \lambda]^T$), and each possible value of $\mathbf{e}$ corresponds to one expert.
(2) each expert only has the most similar experts (i.e., similarity degree $\delta = 1$) as local contacts, thus forming an m-dimensional grid (Figure 3 illustrates a sample two-dimensional grid). As Figure 3 shows, if an expert is not at the grid boundary, then she has $2m$ local contacts; otherwise, the number of local contacts is between m and $2m$.

Then, again based on the above network substrate, long-range contacts are constructed following the inverse rth power distribution for each expert (see the example in Figure 3).

Discussions: Unlike the unified model, in the diversified model, the total abilities across different experts may be different. In particular, the number of experts with the total ability ϕ is $\sum_{q=0}^{\min(m,(\phi-m)/\lambda)}[(-1)^q\binom{m}{q}\binom{\phi-1-q\lambda}{m-1}]$, and the expected value of total ability is $(m + m\sqrt[m]{n})/2$. By these numerical expressions, the distribution of total abilities is reported in Figure 4 under different values of m. The most important property of the diversified model revealed by Figure 4 is that the expert total ability follows a Gaussian-like distribution as m increases. Therefore, the diversified model (when $m > 2$) represents a real expert network that is abundant in experts with average problem solving abilities, while lacks experts with significantly superior/inferior total abilities [8].

Extension of the Unified Model to $m > 2$: There are many strategies to extend the unified model to the case of $m > 2$. Here, we introduce one intuitive method; other methods are left for future work. Specifically, when $m = 2$, each row in the unified model essentially corresponds to the nodes in the diagonal dotted circle in the diversified model (see Figure 3). Therefore, one way to extend the unified model to the case of $m = 3$ is replacing each row in the unified model by the diagonal two-dimensional plane in the diversified model (marked by color green in Figure 5). Hence, for the case of $m > 3$, each row in the unified model can be replaced by an $(m - 1)$-dimensional plane, where all experts have the same total abilities. Detailed analysis under such model extensions can be performed based on the results in this article, thus omitted due to space limitations.

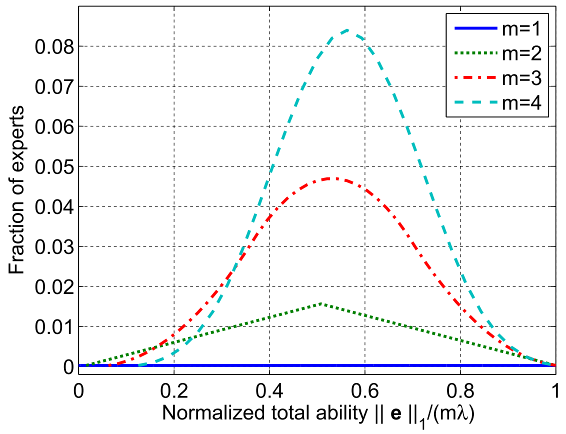


Fig. 4. Total ability distribution in the diversified model ($n = 4096$).

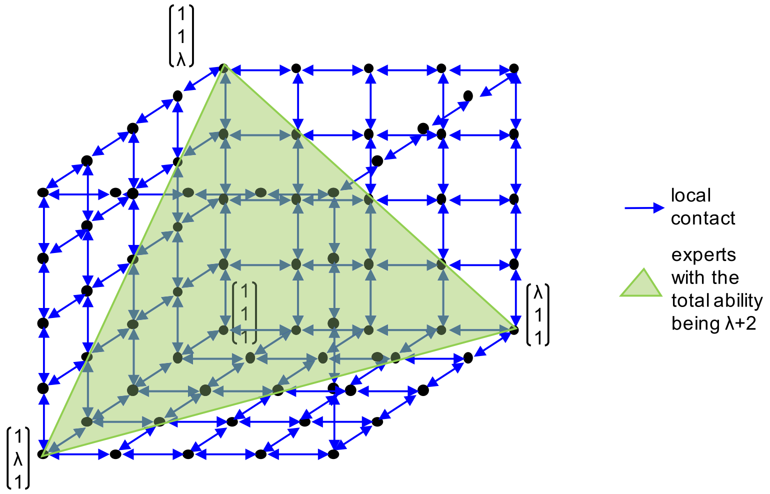


Fig. 5. Diversified model ($\delta = 1$, $m = 3$, $[\cdot]$: expertise vector).

Remark: The theoretical results in this article are based on the network models proposed in this section, where specific expertise distributions over experts are required. Nevertheless, we point out that even if such requirement is not strictly satisfied, our theoretical results still demonstrate high accuracy in predicting query routing performance (see the case study of real datasets in Section 6), thus making contributions from both theoretical and practical perspectives.

4 EFFICIENCY OF DECENTRALIZED SEARCH IN EXPERT NETWORKS

Recall that we assume (in Section 2) that the problem area and the difficulty level in queries are generated uniformly (up to the maximum problem solving ability per area in the expert network) at random, and the first query holder is also arbitrarily chosen. We now present the corresponding quantitative performance bounds and analysis[2] of decentralized search under the network models

[2] All these theoretical results are investigated under the assumption that all query difficulties and the first query holders are uniformly distributed; for other distributions, they are left for future work.

proposed in Section 3. We then discuss how the small-world phenomenon is related to decentralized search in expert networks. Complete theoretical proofs are presented in Section 5.

4.1 Statement of Main Results

Under the unified and diversified models, we have the performance bounds of decentralized search stated as below, where $\ln(\cdot)$ denotes natural logarithm. In these results, we assume that n is sufficiently large such that $n > 8h$ in the unified model, $\sqrt[m]{n} > 8$ in the diversified model, and $n \gg k$ in both models (these assumptions are used to avoid the trivial cases where the first query holders are already very close to the destinations even without long-range contacts; see the theorem proofs in Section 5 for details).

Theorem 4.1. *The average routing path length generated by decentralized search is monotonically increasing with r in both unified and diversified models.*

Theorem 4.2. *The average routing path length generated by decentralized search in the unified model is upper bounded by*

$$O\left(\left(\ln \frac{n}{h}\right)^{r+1}\right) \quad \textit{for } 0 \leq r \leq 1.$$

Theorem 4.3. *The average routing path length generated by decentralized search in the unified model is lower bounded by*

$$\Omega\left(\frac{1}{k^{1/r}} \cdot \left(\frac{n}{h}\right)^{\frac{r-1}{r}}\right) \quad \textit{for } r > 1.$$

Theorem 4.4. *The average routing path length generated by decentralized search in the diversified model is upper bounded by*

$$O\left(\frac{1}{m^r} \cdot (\ln n)^{r+1}\right) \quad \textit{for } 0 \leq r \leq 1.$$

Theorem 4.5. *The average routing path length generated by decentralized search in the diversified model is lower bounded by*

$$\Omega\left(\frac{1}{k^{1/r}} \cdot n^{\frac{r-1}{mr}}\right) \quad \textit{for } r > 1.$$

Corollary 4.6. *Any routing path length generated by decentralized search is upper bounded by n/h in the unified model, and $\sqrt[m]{n}$ in the diversified model.*

4.2 Performance of Decentralized Search in Expert Networks

4.2.1 $0 \leq r \leq 1$. Theorems 4.2 and 4.4 show that when $0 \leq r \leq 1$, the average routing path length using decentralized search is upper bounded by a polylogarithmic function, i.e., a polynomial function of $\ln n$. Therefore, decentralized search is highly efficient in expert searching when $0 \leq r \leq 1$. Such high efficiency occurs mainly for the following two reasons: (i) In expert networks, experts exhibit a certain level of randomness in inter-connections, which causes the formation of a network gradient that drives the query to the destination via decentralized search. (ii) Qualified experts for a given query may not be unique, i.e., query routing terminates at any expert who is capable of resolving this query; therefore, the query routing problem in expert networks is an *anycast* problem. Hence, if a query is routed to an over-qualified expert, then this query does not need to be further routed to the expert with the exact required knowledge level as the problem is already solvable. Another significant insight revealed by Theorems 4.2 and 4.4 is that decentralized search is most effective when $r = 0$ (i.e., long-range contacts are randomly selected) as the number of experts n is sufficiently large according to our assumption (i.e., $\ln(n/h) > 1$ in the unified model and $(\ln n)/m > 1$ in the diversified model); the corresponding routing path length is only logarithmic in the network size.

4.2.2 $r > 1$. When r increases, Theorem 4.1 shows that the average routing path length also rises. When $r > 1$, as shown in Theorems 4.3 and 4.5, the average routing path length can no longer be expressed as a polylogarithmic function, which indicates the ineffectiveness of decentralized search. Nevertheless, Corollary 4.6 proves that for any values of r, there is always an upper bound, which is determined by the network size. Therefore, the worst performance of decentralized search happens when r approaches $+\infty$, for which the lower bound is $\Omega(n/h)$ for the unified model, and $\Omega(\sqrt[m]{n})$ for the diversified model. Comparing to Corollary 4.6, this result suggests that the performance bounds in Theorems 4.3 and 4.5 are tight when r is large. On the other hand, Theorems 4.3 and 4.5 also show that although a larger k may increase the probability of connecting to qualified experts, the impact of k in reducing the average routing path length is weakened when $r > 1$. This is because when r is large, long-range contacts tend to only exist in the vicinity of each expert or even overlap with local contacts, and thus the contribution of a large k diminishes.

4.3 Uniqueness of Small-World Phenomenon in Expert Networks

As shown in [4, 5, 21], the basic standard for justifying networks with the small-world phenomenon is that the associated network diameters can be expressed as a polylogarithmic function of n. Thus, by Theorems 4.2 and 4.4, we conclude that not only does the small-world phenomenon *exist* in expert networks under *both* unified and diversified models when $0 \le r \le 1$, but also decentralized search is able to find these short paths. This is in sharp contrast with the unicast problem in prior works on the small-world phenomenon. Specifically, assuming that individual connections also follow the inverse rth power distribution (constructed based on their lattice distances) in the unicast problem, the work in [20] shows that though the small-world phenomenon is pervasive for a range of r, decentralized search is *only* efficient under a unique value of r. This is because in a unicast problem with the destination being t, along the way from the current message holder to t, if one intermediate node routes this message to a long-distant node (corresponding to one over-qualified expert in expert networks) that is beyond t, then this message needs to be routed back to t, thus resulting in longer routing paths.

5 THEORETICAL PROOFS

To prove the theorems in Section 4, let (i, τ) denote a query posted to the network, and s the first holder of (i, τ). We define L, a random variable, as the total number of hops spent on finding a qualified expert for (i, τ) (i.e., starting from s and terminating at any expert that can solve (i, τ)). Then, it suffices to determine the expected value $\mathbb{E}[L]$ under different network settings.

5.1 Proof of Theorem 4.1

Regarding query (i, τ), there exists at least one qualified expert according to our assumption. Let T be the set of all experts qualified to solve (i, τ). We sort all experts in T in an increasing order w.r.t. their expertise in area i such that $e_i^{(t_1)} \le e_i^{(t_2)} \le \cdots \le e_i^{(t_\zeta)}$, where $T = \{t_1, t_2, \ldots, t_\zeta\}$. For all other experts, let U denote the set containing all experts with expertise better than s but less than t_1 in area i. Then, when delivering (i, τ) from s to T, only experts in U can become the relay experts. Furthermore, from s to T, at each relay expert w, $e_i^{(t_1)} - e_i^{(w)}$ is strictly decreasing, as each expert in both unified and diversified models has at least one (local or long-range) contact who has better expertise in area i. Therefore, we can divide this routing process (before reaching any expert in T)

into several phases, where in phase j, $e_i^{(t_1)} - e_i^{(w)}$ w.r.t. the current query holder w is greater than $j\Delta$ and at most $(j+1)\Delta$ (Δ is a constant). Then, suppose the query routing process is currently in phase j, then the probability of leaving phase j in the next routing step is proportional to the probability of experts in phase j having long-range contacts to experts in phase q ($q < j$) or experts in T (note that the normalization factor in the inverse rth power distribution w.r.t. an expert is fixed as it only relies on the underlying network structure). Moreover, the probability of having such long-range contacts is strictly increasing with $1/r$ according to the inverse rth power distribution. Thus, if experts associated with phase j are employed as relay experts, then the average number of hops in phase j is monotonically increasing with r. We note that by the definition of phase, the number of hops in certain phases can be zero. Nevertheless, the probability of experiencing a certain phase, say phase j, is also strictly increasing with r. Thus, the overall average number of hops in phase j remains monotonically increasing with r. As $\mathbb{E}[L|(i,\tau)]$ for the given query (i,τ) is the sum of average number of hops in each phase, we have that $\mathbb{E}[L|(i,\tau)]$ is monotonically increasing with r. The above argument remains valid regarding all other queries, we therefore conclude that the total expectation $\mathbb{E}[L]$ is monotonically increasing with r for both the unified and diversified model. □

5.2 Proof of Theorem 4.2

To prove Theorem 4.2, we first prove that the following inequality holds:

$$\frac{d^{-r}}{\sum_{j=1}^{n'} j^{-r}} \geq \frac{1}{n'(\sum_{j=1}^{n'} j^{-1})^r} \text{ for } 0 \leq r \leq 1, \text{ and } 1 \leq d \leq n'. \tag{1}$$

Let $g(x) := x^r$, where $0 \leq r \leq 1$. Define random variable y as follows: $y = 1/j$ with probability $1/n'$ for $j = 1, 2, \ldots, n'$. Then, we have

$$g(\mathbb{E}[y]) = \left(\frac{\sum_{j=1}^{n'} j^{-1}}{n'}\right)^r, \ \mathbb{E}[g(y)] = \frac{\sum_{j=1}^{n'} j^{-r}}{n'}.$$

Since $g(x)$ is a concave function, by Jensen's theorem [6], the inequality $g[\mathbb{E}(x)] \geq \mathbb{E}[g(x)]$ holds. Therefore,

$$n'\left(\sum_{j=1}^{n'} j^{-1}\right)^r \geq (n')^r \sum_{j=1}^{n'} j^{-r} \geq d^r \sum_{j=1}^{n'} j^{-r},$$

as $1 \leq d \leq n'$. Thus, (1) is correct.

Now, we can prove Theorem 4.2. In the unified model, starting from one expert u, there are at most $2h$ experts with the expertise distance being j (j is an integer). Thus, we have

$$\sum_{w \in C_u} [d(u \rightarrow w)]^{-r} \leq \sum_{j=1}^{(n/h)-1} 2hj^{-r}. \tag{2}$$

According to the distribution of long-range contacts, $\Pr(u \rightarrow w) = [d(u \rightarrow w)]^{-r} / \sum_{v \in C_u} [d(u \rightarrow v)]^{-r}$; therefore, by (2),

$$\Pr(u \rightarrow w) \geq \frac{[d(u \rightarrow w)]^{-r}}{\sum_{j=1}^{(n/h)-1} 2hj^{-r}}. \tag{3}$$

Since $d(u \to w) \leq (n/h) - 1$ and $0 \leq r \leq 1$, by (1), (3) is further lower bounded by

$$\frac{[d(u \to w)]^{-r}}{\sum_{j=1}^{(n/h)-1} 2hj^{-r}} \geq \frac{1}{2h\left(\frac{n}{h} - 1\right)\left(\sum_{j=1}^{(n/h)-1} j^{-1}\right)^r} \geq \frac{1}{2(n-h)\left(1 + \ln\frac{n}{h}\right)^r} \geq \frac{1}{2(n-h)\left(\ln\frac{3n}{h}\right)^r}.$$

Thus, $\Pr(u \to w) \geq [2(n-h)(\ln\frac{3n}{h})^r]^{-1}$. For the received query (i, τ), suppose there are η qualified experts in each row of the unified model. Then, together there are ηh experts, denoted by set T, capable of solving (i, τ). Let $\Pr(u \to T)$ denote the probability that u has a contact in T when there are k out-going edges, then $\Pr(u \to T) \geq \sum_{w \in T} \Pr(u \to w) \geq \eta h [2(n-h)(\ln\frac{3n}{h})^r]^{-1}$. Let $\Gamma := \eta h[2(n-h)(\ln\frac{3n}{h})^r]^{-1}$ and Y_η denote the total number of hops that are spent for solving (i, τ) when there are ηh qualified experts. We have

$$\mathbb{E}[Y_\eta] = \sum_{j=1}^{\infty} \Pr[Y_\eta \geq j] \leq \sum_{j=1}^{\infty} (1-\Gamma)^{j-1} = \frac{1}{\Gamma}. \tag{4}$$

Recall that L denotes the total number of hops spent for solving (i, τ). Since i and τ in (i, τ) are uniformly distributed ($i = \{1, \ldots, m\}$, $0 < \tau \leq (n/h) - 1$), we can derive

$$\begin{aligned}\mathbb{E}[L] &= \sum_{\eta=1}^{(n/h)-1} \frac{1}{(n/h)-1} \mathbb{E}[Y_\eta] \\ &\leq \frac{2(n-h)\left(\ln\frac{3n}{h}\right)^r}{h} \cdot \frac{h}{n-h} \cdot \sum_{\eta=1}^{(n/h)-1} \eta^{-1} \\ &\leq 2\left(\ln\frac{3n}{h}\right)^r \left(1 + \ln\frac{n}{h}\right) \leq 2\left(\ln\frac{3n}{h}\right)^{r+1}.\end{aligned} \tag{5}$$

Therefore, the average routing path length $\mathbb{E}[L]$ is upper bounded by $O((\ln\frac{n}{h})^{r+1})$, when $0 \leq r \leq 1$. □

5.3 Proof of Theorem 4.3

In the unified model, for any expert u, $\sum_{w \in C_u} [d(u \to w)]^{-r}$ is lower bounded by a constant ξ (ξ is the cardinality of set $S := \{w \in V \setminus \{u\} : d(u \to w) = 1, e_i^{(w)} - e_i^{(u)} = 1\}$). Then, the probability that u has a long-range contact w with $e_i^{(w)} - e_i^{(u)}$ greater than l, denoted by $\Pr[e_i^{(w)} - e_i^{(u)} > l]$, is

$$\Pr[e_i^{(w)} - e_i^{(u)} > l] \leq \frac{\sum_{j=l+1}^{(n/h)-1} \xi j^{-r}}{\xi} \leq \int_l^\infty x^{-r} dx = \frac{l^{1-r}}{r-1}. \tag{6}$$

Then, following similar arguments in [20], we define $\alpha := 1/r$, $\beta := (r-1)/r$, and $\theta := \min(r-1, 1)/(8k)$. Let T denote the set of all experts who can solve (i, τ) ($|T| \geq 1$). We define $\mathcal{A}_j$ to be the event that in the jth hop since decentralized search starts query routing, query (i, τ) reaches an expert w ($w \notin T$) that has a long-range contact v with $e_i^{(v)} - e_i^{(w)} \geq (kn/h)^\alpha$. Then, let $\mathcal{A} := \bigcup_{1 \leq j \leq \theta(kn/h)^\beta} \mathcal{A}_j$ denote the event that this occurs in the first $\theta(kn/h)^\beta$ hops. As the probability of a union of events is upper bounded by the sum of their individual probabilities, we have

$$\Pr[\mathcal{A}] \leq \sum_{1 \leq j \leq \theta(kn/h)^\beta} \Pr[\mathcal{A}_j] \leq \theta(kn/h)^\beta \frac{k(kn/h)^{\alpha - \alpha r}}{r-1} \leq \frac{1}{8}. \tag{7}$$

Let t_1 be the expert in T with the minimum expertise in solving (i, τ). Define $\mathcal{B}$ to be the event that $e_i^{(t_1)} - e_i^{(s)} > \frac{n}{8h}$ (recall that s denotes the first query holder). Since $n/h > 8$ according to our assumption, we have

$$\Pr[\mathcal{B}] = \sum_{x=1}^{\frac{n}{h} - \frac{n}{8h}} \frac{1}{(n/h) - 1} \cdot \frac{\frac{n}{h} - x - \frac{n}{8h}}{n/h} > 1/3.$$

By (7), $\Pr[\overline{\mathcal{B}} \vee \mathcal{A}] < \frac{2}{3} + \frac{1}{8}$; therefore, $\Pr[\mathcal{B} \wedge \overline{\mathcal{A}}] > \frac{5}{24}$.

We now prove that if $\mathcal{B}$ occurs and $\mathcal{A}$ does not occur, then $\mathbb{E}[L|\mathcal{B} \wedge \overline{\mathcal{A}}] \geq \theta(kn/h)^{\beta}$ as follows: If $\mathcal{A}$ does not occur, then at each routing step, query (i, τ) can move toward T by a distance of at most $(kn/h)^{\alpha}$ in area i; then after $\theta(kn/h)^{\beta}$ steps, the total moved distance in area i is at most

$$\theta(kn/h)^{\alpha+\beta} = \theta k(n/h) = \min(r-1, 1)n/(8h) \leq \frac{n}{8h},$$

i.e., (i, τ) cannot reach any expert in T if $\mathcal{B}$ occurs and $\mathcal{A}$ does not occur. Therefore, $\mathbb{E}[L|\mathcal{B} \wedge \overline{\mathcal{A}}] \geq \theta(kn/h)^{\beta}$. Accordingly,

$$\begin{aligned} \mathbb{E}[L] &\geq \mathbb{E}[L|\mathcal{B} \wedge \overline{\mathcal{A}}] \cdot \Pr[\mathcal{B} \wedge \overline{\mathcal{A}}] \\ &> \frac{5}{24}\theta(kn/h)^{\beta} = \frac{5\min(r-1, 1) \cdot k^{-\frac{1}{r}}}{192} \left(\frac{n}{h}\right)^{\frac{r-1}{r}}. \end{aligned}$$

Therefore, when $r > 1$, the average routing path length in the unified model is lower bounded by $\Omega(k^{-\frac{1}{r}}(\frac{n}{h})^{\frac{r-1}{r}})$. □

5.4 Proof of Theorem 4.4

We follow similar arguments as those in the proof of Theorem 4.2. In the diversified model, there are up to $c_0\lambda^{m-1}$ (c_0: a constant) experts with the same expertise distance from any expert; therefore, under the diversified model, (2) in the unified model is changed to $\sum_{w \in C_u}[d(u \to w)]^{-r} \leq \sum_{j=1}^{m\lambda - m} c_0\lambda^{m-1}j^{-r}$ in the diversified model. Since $d(u \to w) \leq m\lambda - m$, (3) becomes

$$\begin{aligned} \Pr(u \to w) &\geq \frac{[d(u \to w)]^{-r}}{\sum_{j=1}^{m\lambda-m} c_0\lambda^{m-1}j^{-r}} \\ \text{by (1)}, \ &\geq \frac{1}{mc_0\lambda^m \left(\sum_{j=1}^{m\lambda-m} j^{-1}\right)^r} \geq \frac{1}{mc_0\lambda^m(\ln(3m\lambda))^r}. \end{aligned}$$

To compute $\mathbb{E}[L]$ for query (i, τ), let $\eta := \lambda - \tau + 1$. Then, there are $\eta\lambda^{m-1}$ qualified experts, denoted by set T. Let $\Pr(u \to T)$ be the probability that u has a long-range contact in T, then $\Pr(u \to T) \geq \eta\lambda^{m-1}\sum_{w \in T}\Pr(u \to w) \geq \eta(c_0m\lambda)^{-1}(\ln(3m\lambda))^{-r}$.Then, following similar arguments for computing (4–5) in the unified model, we have $\mathbb{E}[L] \leq c_0m(\ln(3m\lambda))^{r+1}$. Thus, when $0 \leq r \leq 1$, $\mathbb{E}[L]$ is upper bounded by $O(m^{-r}(\ln n)^{r+1})$, where we use $\lambda^m = n$. □

5.5 Proof of Theorem 4.5

Let T denote the set of all experts who can solve query (i, τ). Similar to the proof of Theorem 4.3, in the diversified model, $\sum_{w \in C_u}[d(u \to w)]^{-r}$ is lower bounded by a constant ξ (defined in Section 5.3); therefore, similar to (6), we have $\Pr[e_i^{(w)} - e_i^{(u)} > l] \leq (\sum_{j=l+1}^{m\lambda-m} \xi j^{-r})/\xi \leq l^{1-r}/(r-1)$. Again, define $\alpha := \frac{1}{r}$, $\beta := \frac{r-1}{r}$, and $\theta := \frac{\min(r-1,1)}{8k}$. Define $\mathcal{A}_j$ to be the event that in the jth hop, query (i, τ) reaches an expert w ($w \notin T$) that has a long-range contact v with $e_i^{(v)} - e_i^{(w)} \geq (k\lambda)^{\alpha}$.

Then, let $\mathcal{A} := \bigcup_{1 \le j \le \theta(k\lambda)^\beta} \mathcal{A}_j$ denote the event that this occurs in the first $\theta(k\lambda)^\beta$ hops. Thus, $\Pr[\mathcal{A}] \le \sum_{1 \le j \le \theta(k\lambda)^\beta} \Pr[\mathcal{A}_j] \le 1/8$.

Next, similar to the proof of Theorem 4.3, let t_1 be the expert in T with the minimum expertise in solving (i, τ), and $\mathcal{B}$ the event that $e_i^{(t_1)} - e_i^{(s)} > \frac{\lambda}{8}$ (recall s is the first query holder). Since $\lambda = \sqrt[m]{n} > 8$ according to our assumption, we again have $\Pr[\mathcal{B}] > \frac{1}{3}$ and $\Pr[\mathcal{B} \wedge \overline{\mathcal{A}}] > \frac{5}{24}$. Then, using similar arguments for proving Theorem 4.3, we have $\mathbb{E}[L|\mathcal{B} \wedge \overline{\mathcal{A}}] \ge \theta(k\lambda)^\beta$. Hence, $\mathbb{E}[L] \ge \mathbb{E}[L|\mathcal{B} \wedge \overline{\mathcal{A}}] \cdot \Pr[\mathcal{B} \wedge \overline{\mathcal{A}}] > \rho k^{-\frac{1}{r}} \lambda^{\frac{r-1}{r}}$, where $\rho := 5\min(r-1, 1)/192$. Since $\lambda = \sqrt[m]{n}$, the lower bound in the diversified model is $\Omega(k^{-\frac{1}{r}} n^{\frac{r-1}{mr}})$ when $r > 1$. □

5.6 Proof of Corollary 4.6

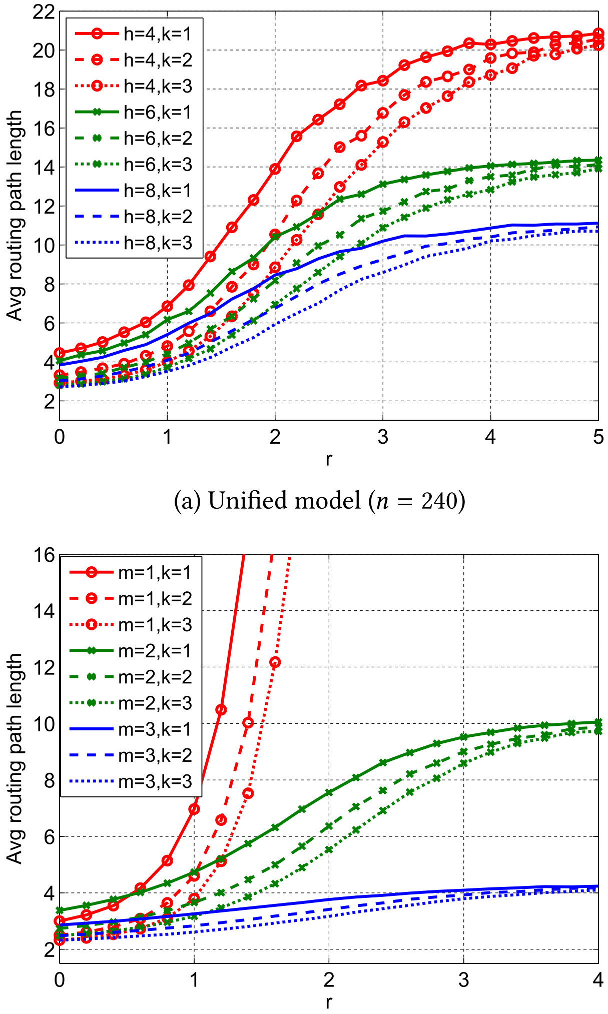
The statement in Corollary 4.6 follows by considering the worst routing case: Query (i, τ) is of the maximum difficulty level (i.e., $\tau = \max_{u \in V} e_i^{(u)}$), and the expert with the worst expertise in area i is selected as the first query holder. □

6 EXPERIMENTS

In this section, we evaluate the performance of decentralized search by computing its average routing path length under the unified/diversified model and explaining the corresponding observations using the performance bounds in Section 4. Then, we study the robustness of decentralized search against query interpretation errors. Finally, we compare the predicted query routing time using the performance bounds to the actual routing time in a case study of real datasets for justifying the applicability of the performance bounds to real networks.

Performance Under the Unified/Diversified Model: To evaluate the performance of decentralized search, we select $n = 240$ and $h = 4, 6, 8$ for the unified model, and $n = 729$ and $m = 1, 2, 3$ for the diversified model, both under $k = 1, 2, 3$. We generate queries by randomly selecting the corresponding problem area and the required expertise level (subject to the network-wise maximum problem solving capability) in both the unified and diversified model; moreover, the first query holders are also randomly chosen. For each network parameter setting, 100 random network realizations are generated, and 500 Monte Carlo runs (each run corresponds to a newly generated query that is randomly distributed to an expert as the first query holder) are conducted on each network realization. Using decentralized search, the resulting routing path length averaged over all network realizations and Monte Carlo runs are reported in Figure 6(a) (unified model) and Figure 6(b) (diversified model).

In Figure 6, as expected, we first observe that the average routing path length increases with r (supported by Theorem 4.1). The most significant conclusion we can draw from Figure 6 is that they confirm the high efficiency of decentralized search when $0 \le r \le 1$ for both network models (as proved in Theorems 4.2 and 4.4). Specifically, compared to the network size (240 and 729 experts in the unified and diversified models), decentralized search achieves extremely small routing path length, i.e., between 2 to 5 when $r = 0$ and 3 to 8 when $r = 1$. Moreover, Figure 6 demonstrates that the performance of decentralized search is similar under different network parameters/models when r is small (e.g., around 0). However, when r increases, the performance under different parameters begins to depart, and the performance under small h (or m) degrades significantly when $r > 1$, for which Theorems 4.3 and 4.5 provide quantitative bounds to capture such performance deterioration. Nevertheless, such performance degradation converges when r is large, because all routing path lengths are constrained by the upper bound (independent of r) established in Corollary 4.6. Moreover, Figure 6 shows that the benefit of a larger number of long-range contacts is not obvious, especially in the case when r is large. This is because, as shown in Theorems 4.3



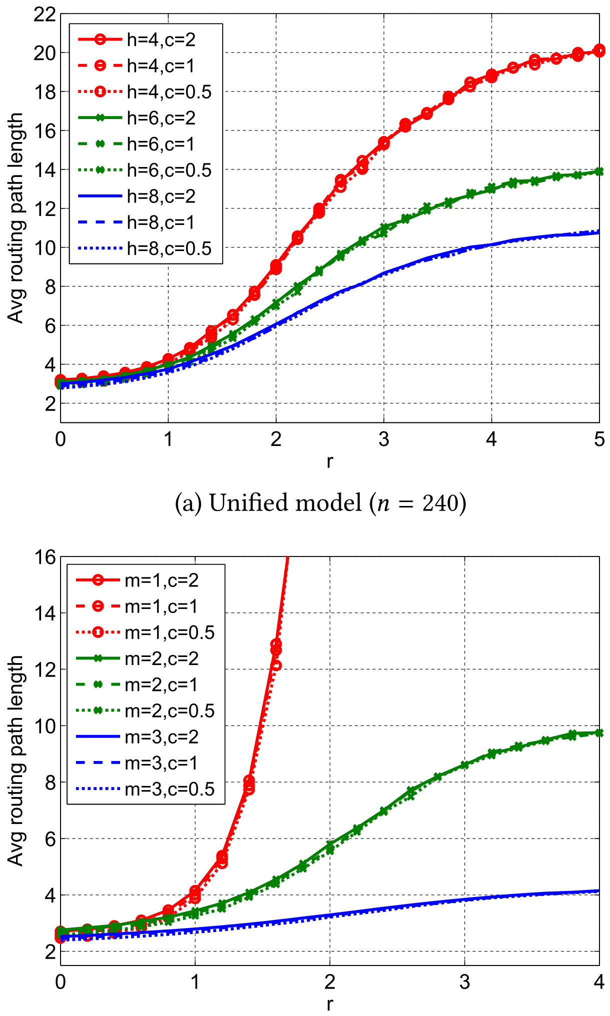
(a) Unified model ($n = 240$)

(b) Diversified model ($n = 729$)

Fig. 6. Average routing path length (100 random network realizations for each network parameter setting, 500 Monte Carlo runs per network realization).

and 4.5, for each expert, having a single contact with significantly different specialties (controlled by r) is more effective than having multiple contacts (controlled by k) with only limited expertise dissimilarity in reducing the average routing path length.

Robustness of Decentralized Search: Thus far, we assume that each expert has accurate estimation of problem difficulties; in other words, the expertise levels required by the queries in different areas are perceived exactly the same as the ground truth for all experts in the network. However, this may not always be true. Therefore, we study how the performance of decentralized search is affected by misinterpreting query difficulties as follows: When attempting to solve query (i, τ), if expert u has sufficient expertise in area i (i.e., $e_i^{(u)} \geq \tau$), then she solves (i, τ); otherwise, her estimation of the expertise required for solving (i, τ) follows the truncated Gaussian distribution with minimum value $\tau_{\min}$, mean μ, and standard deviation σ being functions of u and (i, τ). In particular,



(a) Unified model ($n = 240$)

(b) Diversified model ($n = 729$)

Fig. 7. Average routing path length with query interpretation errors ($k = 3$, 100 random network realizations for each network parameter setting, 500 Monte Carlo runs per network realization).

$\tau_{\min} =: e_i^{(u)}$, $\mu := \tau$, and $\sigma := c(\tau - e_i^{(u)})$, where c is a scaling factor. She then uses this estimated value τ' to find the best next hop from her contacts following the decentralized search rule. In this estimation error model, we capture the fact that in real networks, query estimation accuracy increases as the expert expertise gets closer to the actual requirement. The results of decentralized search under different levels (controlled by the scaling factor c) of such estimation errors averaged over multiple Monte Carlo runs are shown in Figure 7(a) (unified model) and Figure 7(b) (diversified model). These results confirm that under query estimation errors, the performance of decentralized search remains stable; therefore, decentralized search is a robust and reliable solution for query routing in expert networks.

Case Study of Real Datasets: Next, using our theoretical results, we analyze real-world query routing data collected from the IT service department of one of Fortune 500 companies throughout

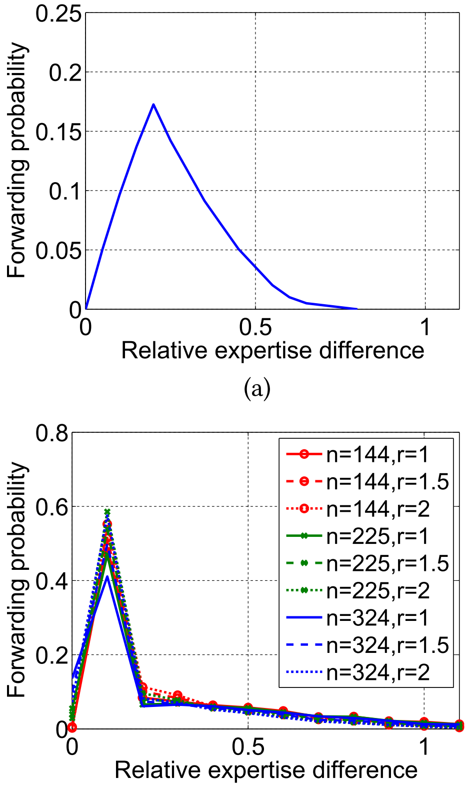
Table 2. Performance in Real Expert Networks (Captured by the Diversified Model Using Decentralized Search, Where $m = 2$ and $k = 1$. $\overline{L}_i$: Average Query Routing Path Length under r_i, $i = 1, 2$)

Datasets	n	r_1	r_2	Real $\overline{L}_1/\overline{L}_2$	Predicted $\overline{L}_1/\overline{L}_2$
OS-1	184	1.98	1.44	1.44	1.63
OS-2	122	1.24	1.04	1.31	1.45
Database	305	1.61	1.01	2.37	2.87
Web service	266	1.14	1.04	1.19	1.26

2006. Depending on query contents, these datasets are categorized into four independent classes: Operating System 1 (OS-1), Operating System 2 (OS-2), Database, and Web Service. Table 2 lists the network parameters in these datasets. For these datasets, we first observe that their network structures can be characterized by the diversified model for the case of $m = 2$ and $k = 1$. In Table 2, two values of r (i.e., r_1 and r_2) are derived from these datasets based on expert inter-connections, where r_2 corresponds to new connections after applying a *mentoring program* to the original expert networks (associated with r_1). In this mentoring program, some less-skilled experts are mentored by experienced experts, which equivalently reduces the value of r. Note that in Table 2, the expert network associated with each dataset is not as strictly structured as the grid structure in Figure 3, i.e., some experts in Figure 3 may be missing; however, the performance prediction remains accurate (see Table 2 for details). In order to apply our theoretical results in Section 4, we still need to justify if the query routing behaviors in these real datasets share any similarities with decentralized search. To this end, we define relative expertise difference as $||\mathbf{e}^{(w)} - \mathbf{e}^{(u)}||_1/||\mathbf{e}^{(u)}||_1$, where w is the next hop expert selected by expert u. The query forwarding probability versus relative expertise difference averaged over all queries received in the four networks of the datasets is shown in Figure 8(a). As comparison, we also compute the same metric based on decentralized search in the diversified model as reported in Figure 8(b), where three networks of similar network sizes and similar values of r_1 and r_2 as those in the real datasets are evaluated. We note that Figure 8(a) and (b) has similar shapes, i.e., the expert with neither too similar nor too different expertise is selected with high probability as the next hop; therefore, the routing behaviors in these datasets do exhibit a certain level of decentralized search. Hence, we can use our results in Section 4 on decentralized search to predict the routing performance in these real datasets. Let $\overline{L}_i$ denote the average query routing path length under r_i ($i = 1, 2$). We compare the real $\overline{L}_1/\overline{L}_2$ with the predicted $\overline{L}_1/\overline{L}_2$ using Theorem 4.5 (as $r_1, r_2 > 1$ for all datasets). The comparison in Table 2 shows that using the theoretical performance bounds, the predicted routing path length is accurate (the error is 21.1% for Database, and 5.9~13.2% for other datasets); therefore, the theoretical results in this article can naturally serve as an efficient tool for analyzing/predicting behaviors in real expert networks. Moreover, these datasets also suggest that to achieve high routing efficiency, network owners can take proactive actions to adjust the expert connections such that the resulting network condition (r_2 is close to 1) approaches the high efficiency region (i.e., $0 \leq r \leq 1$).

7 CONCLUSION

We investigated the efficiency of local-information-based decentralized search for query answering in expert networks, focusing on quantifying the performance of decentralized search under various network settings. Incorporating common expert social inter-connection tendencies, we proposed two expert network models, each representing a unique distribution of expert problem solving abilities in the network. Under these two network models, we established fundamental theories demonstrating when decentralized search is exceptionally effective in finding short query



(a)

(b)

Fig. 8. (a) Query routing behavior in real datasets; (b) query routing behavior of decentralized search in the diversified model ($m = 2$, $k = 1$, 100 random network realizations for each network parameter setting, 500 Monte Carlo runs per network realization).

routing paths. In cases where decentralized search is ineffective, we also quantified how the performance deterioration is correlated to network structures. Evaluations and comparisons of these theoretical results in both synthetic networks and real datasets confirm the efficiency/robustness of decentralized search in expert networks as well as the significance of the developed performance bounds in guiding real network design.

ACKNOWLEDGMENTS

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. 2001. Search in power-law networks. *Physical Review E* 64 (2001), 046135.

[2] Krisztian Balog, Leif Azzopardi, and Maarten De Rijke. 2006. Formal models for expert finding in enterprise corpora. In *ACM SIGIR*.

[3] Arindam Banerjee and Sugato Basu. 2008. A social query model for decentralized search. In *ACM SNAKDD*.

[4] B. Bollobás and F. R. K. Chung. 1988. The diameter of a cycle plus a random matching. *SIAM Journal on Discrete Mathematics* 1 (1988), 328–333.

[5] B. Bollobás and Fernandez De La Vega. 1982. The diameter of random regular graphs. *Combinatorica* 2 (1982), 125–134.

[6] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization.* Cambridge University Press.

[7] Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *ACL.*

[8] Philippe Chassy and Fernand Gobet. 2011. Measuring chess experts' single-use sequence knowledge: An archival study of departure from 'theoretical' openings. In *PloS One.*

[9] Martin Dietzfelbinger and Philipp Woelfel. 2014. Tight lower bounds for greedy routing in higher-dimensional small-world grids. In *ACM-SIAM SODA.*

[10] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. 2017. Metapath2Vec: Scalable representation learning for heterogeneous networks. In *ACM KDD.*

[11] M. Draief and A. Ganesh. 2006. Efficient routeing in Poisson small-world networks. *Journal of Applied Probability* 43 (2006), 678–686.

[12] Alberto Garcia Duran and Mathias Niepert. 2017. Learning graph representations with embedding propagation. In *NIPS.*

[13] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable feature learning for networks. In *ACM KDD.*

[14] John Guare. 1990. *Six Degrees of Separation: A Play.* Vintage Books.

[15] S. Guo, Q. Wang, B. Wang, L. Wang, and L. Guo. 2017. SSE: Semantically smooth embedding for knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering* 29, 4 (2017), 884–897.

[16] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NIPS.*

[17] Xiao Huang, Jundong Li, and Xia Hu. 2017. Label informed attributed network embedding. In *ACM WSDM.*

[18] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR.*

[19] Jon Kleinberg. 2000. Navigation in a small world. *Nature* 406 (2000), 845.

[20] Jon Kleinberg. 2000. The small-world phenomenon: An algorithmic perspective. In *ACM STOC.*

[21] Jon Kleinberg. 2006. Complex networks and decentralized search algorithms. In *International Congress of Mathematicians.*

[22] Jon Kleinberg and Prabhakar Raghavan. 2005. Query incentive networks. In *IEEE FOCS.*

[23] Charles Korte and Stanley Milgram. 1970. Acquaintance networks between racial groups: Application of the small world method. *Journal of Personality and Social Psychology* 15 (1970), 101–108.

[24] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. 2016. Gated graph sequence neural networks. In *ICLR.*

[25] Gengxin Miao, Louise E. Moser, Xifeng Yan, Shu Tao, Yi Chen, and Nikos Anerousis. 2010. Generative models for ticket resolution in expert networks. In *ACM SIGKDD.*

[26] Stanley Milgram. 1967. The small world problem. *Psychology Today* 67 (1967), 61–67.

[27] Annamalai Narayanan, Mahinthan Chandramohan, Lihui Chen, Yang Liu, and Santhoshkumar Saminathan. 2016. Subgraph2Vec: Learning distributed representations of rooted sub-graphs from large graphs. *CoRR* abs/1606.08928 (2016).

[28] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *ICML.*

[29] Pavel Serdyukov, Henning Rode, and Djoerd Hiemstra. 2008. Modeling multi-step relevance propagation for expert finding. In *ACM CIKM.*

[30] Qihong Shao, Yi Chen, Shu Tao, Xifeng Yan, and Nikos Anerousis. 2008. Efficient ticket routing by resolution sequence mining. In *ACM SIGKDD.*

[31] Huan Sun, Mudhakar Srivatsa, Shulong Tan, Yang Li, Lance M. Kaplan, Shu Tao, and Xifeng Yan. 2014. Analyzing expert behaviors in collaborative networks. In *ACM SIGKDD.*

[32] Jeffrey Travers and Stanley Milgram. 1969. An experimental study of the small world problem. *Sociometry* 32 (1969), 425–443.

[33] Zhigang Wang and Juanzi Li. 2016. Text-enhanced representation learning for knowledge graph. In *IJCAI.*

[34] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of "small-world" networks. *Nature* 393 (1998), 440–442.

[35] Han Xiao, Minlie Huang, Lian Meng, and Xiaoyan Zhu. 2017. SSP: Semantic space projection for knowledge graph embedding with text descriptions. In *AAAI.*

[36] Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2016. Representation learning of knowledge graphs with hierarchical types. In *IJCAI.*

[37] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2016. Revisiting semi-supervised learning with graph embeddings. In *ICML.*

[38] Liang Yao, Yin Zhang, Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang, and Qinfei Chen. 2017. Incorporating knowledge graph embeddings into topic modeling. In *AAAI*.

[39] JianYang Zeng and WenJing Hsu. 2006. Optimal routing in a small-world network. *Journal of Computer Science & Technology* 21 (2006), 476–481.

[40] Haoqi Zhang, Eric Horvitz, Yiling Chen, and David C. Parkes. 2012. Task routing for prediction tasks. In *AAMAS*.

Received September 2016; revised October 2018; accepted December 2018