A Social Identity Approach to Identify Familiar Strangers in a Social Network

Nitin Agarwal, Huan Liu, Sudheendra Murthy, Arunabha Sen, and **Xufei Wang**
School of Computing and Informatics
Arizona State University
{Nitin.Agarwal.2, Huan.Liu, sudhi, asen, Xufei.Wang}@asu.edu

Abstract

We present a novel problem of searching for 'familiar strangers' in a social network. Familiar strangers are individuals who are not directly connected but exhibit some similarity. The power-law nature of social networks determines that majority of individuals are directly connected with a small number of fellow individuals, and similar individuals can be largely unknown to each other. Moreover, the individuals of a social network have only a local view of the network, which makes the problem of aggregating these familiar strangers a challenge. In this work, we formulate the problem, show why it is significant to address the challenge, and present an approach that innovatively employs the social identities of the individuals with competitive approaches. The blogger and citation network are used to showcase technical details and empirical results with related issues and future work.

Introduction

Familiar strangers as defined by Stanley Milgram (Milgram 1972) in physical world are those individuals who do not know each other but share some common attributes like interests, occupation, location, etc. For instance, people taking the same train daily find familiar faces but do not know each other. Analogous to the physical world, it is equally interesting and challenging to define and study the existence of familiar strangers in virtual or online world. Social networks represent a complex set of human relations through interactions expressed via a spectrum of social media websites like blogs, online friendship networks, wikis, media sharing websites, social tagging websites and etc. In an online world, familiar strangers could be defined as those individuals who are not friends with each other, i.e., they are not in each others social network, but they share some common set of attributes like hobbies, community affiliations, workplace, location, etc. A more formal definition is given later.

Identifying familiar strangers has profound applications in online social networks. Since the online social networks are shown to have long tail distribution, i.e., most of the members have very few contacts and very few members have a large number of contacts, which means that most of these members do not know each other. Although many of them could have a lot in common but due to the long tail distribution it is quite likely that they may not know each other. Aggregating such familiar strangers could form a critical mass such that (1) the understanding of one member gives us a sensible and representative glimpse to others, (2) more data about familiar members can be collected for better customization and services (e.g., personalization and recommendation), (3) the nuances among them suggest new business opportunities, and (4) knowledge about them can facilitate predictive modeling and trend analysis in new product/market development. Connecting them to form a critical mass can only potentially expand their social network, i.e., job searching, special interest group formation. Aggregating familiar strangers can encourage participation due to the crowd effect (Kumar et al. 2004). People usually trust those with similar interests. Knowledge transfer or information flow among friends and acquaintances becomes smoother and more receptive.

Identifying familiar strangers in online social networks is interesting and involves several key challenges. Individuals have only local view, i.e., individuals know their contacts but may not know their contacts' contacts and so on. Searching for all the contacts of a node, his contacts' contacts and so on, to identify familiar strangers incurs an exponential cost. Each individual is associated with some content or attributes. The challenge lies in intelligently putting that information to the benefit of searching familiar strangers. Evaluation and validation of the proposed approaches is a big issue due to the absence of an established ground truth.

Problem Formulation

Here we define familiar strangers and formulate the problem of searching them using local information. Given a social network $\mathcal{G}$ where V is the set of vertices (nodes) or the members of the social network. The nodes are associated with an attribute. The attribute can take one or more values from a domain $\mathcal{D} = \{a_1, a_2, ..., a_l\}$. We call this the *attribute-value* set of a node and is denoted by A_u for a node u ($u \in V$). Each node u has a local view of the network (also known as an egocentric view (Wasserman and Faust 1994)), that means the node only knows its *adjacent* nodes denoted by $C_u = \{m_1, m_2, ..., m_y \mid edge(u, m_p) \neq 0,\ 1 \leq p \leq y\}$, also known as u's contacts. Here $edge(c, d) \neq 0$ denotes an edge between nodes c and d. This is similar to a scenario where one knows his/her friends but doesn't know

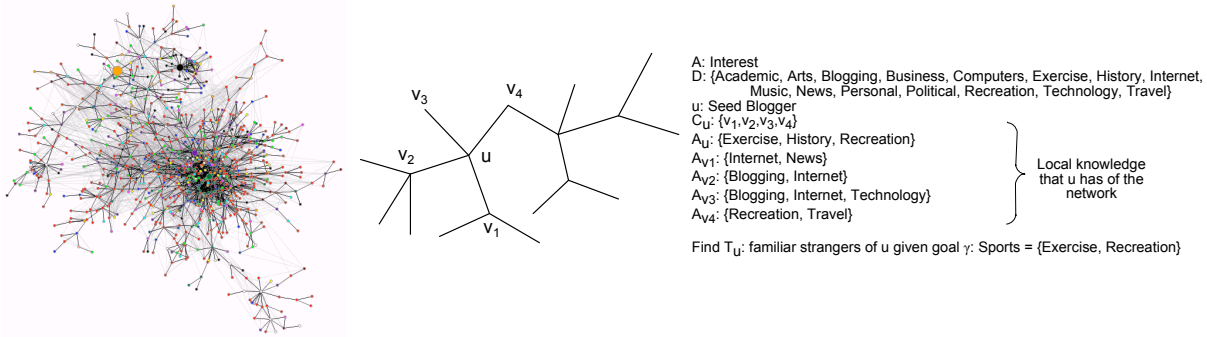

Figure 1: Searching familiar strangers for a node u given the local network information that u has and the goal γ.

his/her friends' friends and so on. In order to define familiar strangers of u, it is essential to define the notion of similarity.

Definition 1 (Similarity) Nodes u and v are similar *iff* $A_v \cap \gamma \neq \emptyset$, where γ is a goal described as $\gamma \subseteq A_u$. ■

Definition 2 (Familiar Strangers) Given u and γ, T_u is the set of familiar strangers of u *iff* (1) for all the nodes $v \in T_u, edge(u, v) = 0$ i.e., all the nodes v are non-adjacent to u - *stranger*[1] and (2) all the nodes v are similar to u with respect to γ as defined above - *familiar*. ■

The problem of searching for familiar strangers given a node u can be illustrated in Figure 1 where a blogger social network is presented in the left, snippet of which is presented in the middle. Here the attribute A is "Interest" and $\mathcal{D}$ is the domain for the values of "Interest". C_u represents the contacts of u and A_u represents the attribute-value set of u. $A_{v_1}, A_{v_2}, A_{v_3}, A_{v_4}$ represent the attribute-value sets of v_1,v_2,v_3, and v_4 respectively. We need to find T_u, familiar strangers of u for the goal γ ("Sports") defined by the combination of "Exercise" and "Recreation".

The challenge lies in searching for familiar strangers efficiently, i.e., in minimum number of edge traversals with local information. To compute the lower bound on the search space for finding the familiar strangers, consider the centralized version of the problem, in which the node u has global or whole view of the network and the objective is to find the smallest set of edges that will connect all the nodes in T_u starting at node u. This centralized version of the familiar strangers problem corresponds to the Steiner tree problem. Given a subset of nodes $V' \subset V$ in a graph $G = (V, E)$, the Steiner tree (T) spans the node set V' with least number of edges. The node set V' is referred to as the *required nodes* or *terminal nodes* and the set of nodes in $V \setminus V'$ is referred to as the *optional nodes* or *Steiner vertices*. It may be noted that tree T contains all the nodes in set V' and zero or more nodes in set $V \setminus V'$. The Steiner tree in a social network that spans the node u and the familiar strangers T_u provides the least number of edges that need to be traversed to find all the familiar strangers of u and thus provides a lower bound on the search space of the familiar strangers problem.

The problem of finding the Steiner tree is known to be NP-complete (Du and Hu. 2008). We provide an Integer Linear Programming (ILP) formulation to solve the Steiner tree problem optimally. Given the *undirected* social network graph $\mathcal{G} = (V, E)$, we first construct the corresponding *directed* graph $H = (V, F)$, in which two directed edges $\{(v_i, v_j), (v_j, v_i)\} \in F$ for each undirected edge $(v_i, v_j) \in E$. Let the number of required nodes be denoted by n, i.e., $|V'| = n$ and let an arbitrary vertex say, the node $u \in V'$ be designated as the root node. The ILP views the directed graph H as a flow graph, in which $(n-1)$ units of flow are routed from the root u towards the nodes in $V' \setminus \{u\}$ through minimum number of edges. Each node in $V' \setminus \{u\}$ consumes exactly one unit of flow. The edges of graph H through which a positive (unit) flow exists form the minimum-edge arborescence[2] in H spanning the vertices V'. The undirected edges in graph G corresponding to the arborescence edges forms the required Steiner tree in G.

Let indicator variables $x_{v_iv_j} = 1$, if edge (v_i, v_j) belongs to the required minimum-edge arborescence T in H, otherwise, $x_{v_iv_j} = 0$. Let variables $f_{v_iv_j} \geq 0$ represent non-negative flow on the edges. The variables $x_{v_iv_j}$ and $f_{v_iv_j}$ are defined for all edges $(v_i, v_j) \in F$. The objective is to minimize the number of edges in the arborescence in H,

$$Minimize \sum_{(v_i,v_j)\in F} x_{v_iv_j}$$

- There are exactly $(n-1)$ units of flow emanating out of the root node u and 0 units of flow going into it. That is,

$$\sum_{(u,v_j)\in F} f_{uv_j} = n-1, \quad \sum_{(v_j,u)\in F} f_{v_ju} = 0$$

- Every other required node, i.e., $v_i \in V' \setminus \{u\}$ consumes 1 unit of flow. That is,

$$\forall v_i \in V' \setminus \{u\}, \sum_{(v_j,v_i)\in F} f_{v_jv_i} - \sum_{(v_i,v_j)\in F} f_{v_iv_j} = 1$$

- A positive flow exists on an edge, *iff* the edge is selected in the arborescence which is ensured by:

$$\forall (v_i, v_j) \in F,\ f_{v_iv_j} \leq (n-1)x_{v_iv_j}$$

Because solving ILP in general takes exponential time, we employ a 2-Approximation algorithm based on Minimum

[1]This definition of stranger nodes is borrowed from the famous concept of *weak ties* (Granovetter 1973).

[2]An arborescence T of a graph H is a directed, rooted tree subgraph of H in which all edges point away from the root.

Spanning Tree approach (Du and Hu. 2008) for computing Steiner trees. The 2-Approximation algorithm produces a solution that is guaranteed to be within 2 times the optimal solution in terms of the edge traversals.

Social Identity Theory

Real-world social networks of people have been shown to exhibit properties of searchability, which means a target can be found quickly even in the absence of global network view (Watts, Dodds, and Newman 2002). Searchability in social networks has been attributed to the tendency of people to cluster their contacts into meaningful groups based on different attributes and selecting relevant cluster of contacts to advance the search at each hop which would take the search closer to the destination. This arrangement of neighbors in groups gives a sense of social identity (Tajfel 1978).

Social identity theory has been widely studied in real-world social networks in terms of observing searchability property of the network. In this paper, we attempt to utilize the social identity theory in online social networks to identify familiar strangers, which is the first of its kind to the best of our knowledge. Directly connected neighbors of a node form the set of its contacts and the attribute-value set of the nodes are used to construct the social identity. More details on social identity construction using the attribute-value set is described in the Social Identity Construction.

Approaches for Egocentric View

Here we present strategies to find familiar strangers T_u of a node u given goal γ using an egocentric view of the network.

Social Identity Approach

According to social identity theory, people cluster their contacts into meaningful groups and pick the cluster that has maximum similarity with the goal γ. So we prune some contacts at each level and propagate the search with the selected cluster of contacts to ensure that the search remains closer to the specified γ instead of wandering away.

Social Identity Construction Social identity based search relies on the ability of a node u to cluster its contacts. Each node in the network is represented as a vector space model of its attributes and simple cosine similarity based measures could be used to compute affinity matrix between contacts of node u. Then conventional clustering algorithms like k-means could be used to cluster the contacts of node u. The clustering approach could be more sophisticated if more data is available about the nodes of the network besides the attribute-value set. For a blogger social network dataset along with the blogger network and their attribute-value set[3] we also have their blog posts and the metadata associated with the blogger like tags, categories, and blog post text. This rich metadata about the bloggers is used to construct the vector space model for each of the contact s of a node u in blogger network. Here the terms of the vector space model are the words in the vocabulary after removing stop words and stemming. However, this is a very sparse and high-dimensional vector. So this sparse vector could be transformed to concept space vector using latent semantic analysis (Deerwester et al. 1990). The transformed vector is less sparse and low dimensional.

Clustering of the contacts could be performed either offline or online while searching. We perform the clustering offline. So the social identity of the nodes of the network are constructed *a priori* to speedup the search process. Online clustering takes advantage of the dynamics of the network, nevertheless, it increases the response time while searching. We can also bypass the construction of social identities to search for familiar strangers. Perhaps this would mean that the search phase will look at all the contacts of a node to find the most relevant nodes to propagate the search. However, by constructing social identity we cluster the contacts and pick the relevant cluster, hence pruning the search space early on. Since clustering is done offline, it does not incur clustering overhead costs while searching.

Example 1: To illustrate with an example, refer to Figure 1, where we need to find the familiar strangers of the node u with respect to the goal, $\gamma = \{\text{Exercise, Recreation}\}$. We can either search all his contacts viz., v_1, v_2, v_3, v_4 to find the contacts that are similar to the γ. This would result in v_4 as the contact whose attribute-values match with the γ. Or we can cluster the contacts offline and pick the relevant cluster. Clustering resulted in two clusters one with v_1, v_2, v_3 and the other with v_4. Now the second cluster with v_4 is more similar to the γ, so we pick the contacts in this cluster, which in this case is v_4. The latter strategy greatly prunes the search space, especially when the nodes have much larger number of contacts[4] and the clustering is performed offline. ■

Social Identity based Search for familiar strangers of a node u and γ can be summarized in the pseudo code in Algorithm 1. Given a node: u, its contacts: C_u, its contacts' attribute-value set: $\mathbf{B}_{C_u}$, and γ as input, it outputs a set of node(s) T_u that are the familiar stranger(s) for u. Algorithm first clusters the contacts C_u of node u and selects the cluster that has maximum similarity with γ, i.e., C'_u. Then among the node(s) in C'_u, node(s) whose attribute-value set matches with γ are selected and we call this set of node(s) C''_u. The node(s) in C''_u are then added to a data structure Q. For each node t in Q search is repeated by first clustering the contacts C_t of node t and then selecting the cluster that has maximum similarity with γ, i.e., C'_t. Then further filter C'_t by selecting the node(s) whose attribute-value set matches with γ. Assign these node(s) to the set C''_t. Node(s) in C''_t are added to the set T_u and Q. Q is a FIFO data structure to ensure a breadth-first search. We do not add C''_u to T_u since these are the adjacent contacts of u and not strangers.

A social network could be a cyclic graph so a person might get multiple requests to search his contacts to find familiar strangers. We assume that a node searches his contacts only once. This is realistic because once a person has

[3]Bloggers' attribute-value set construction is explained in more detail in the BlogCatalog section.

[4]It has been found that on average people have approximately over 150 contacts, also known as the Dunbar number (Bialik 2007).

Input : Node: u,
Contacts of u: $\overrightarrow{C_u}$,
Attribute-value set of contacts of u: $\mathbf{B_{C_u}}$,
A goal: γ
Output: Set of nodes T_u familiar strangers to node u

1 $T_u \leftarrow \emptyset$;
2 Cluster the contacts C_u of node u;
3 $C'_u \leftarrow$ select the cluster of contacts that has maximum similarity with γ;
4 $C''_u \leftarrow$ select the nodes from C'_u whose attribute-value set match with γ;
5 Add selected nodes C''_u to a FIFO data structure Q;
6 Set *participatedFlag* for $u \leftarrow true$;
7 **while** $Q \neq \emptyset$ **do**
8 $t \leftarrow$ dequeue a node from Q;
9 **if** *participatedFlag for node* $t = false$ **then**
10 $C_t \leftarrow$ All contacts of t;
11 Cluster the contacts C_t of node t;
12 $C'_t \leftarrow$ select the cluster of contacts that has maximum similarity with γ;
13 $C''_t \leftarrow$ select the nodes from C'_t whose attribute-value set match with γ;
14 Add selected nodes C''_t to Q;
15 Add C''_t to T_u;
16 Set *participatedFlag* for node $t \leftarrow true$;
17 **end**
18 **end**
19 Return T_u;

Algorithm 1: Searching familiar strangers of u.

searched his contacts and forwarded the search request to his contacts he has no incentive to do it again. This is realized by associating a `participatedFlag` to each node which is set to false by default and is set to true once the node gets a search request and forwards it to his contacts (line 16 in the Algorithm 1). A node checks the `participatedFlag` before searching its contacts and propagating the search to the next hop (line 9 in the Algorithm 1).

Exhaustive Search Approach

Here a node explores all his contacts and his contacts explores all their contacts and so on to search for the nodes that have maximum similarity with the goal γ. This procedure continues till all the familiar strangers T_u of the node u are found. This exhaustive search procedure incurs an exponential computational cost. Approximately, for an average degree d of the network, and h hops needed to find all the familiar strangers, the exhaustive approach needs to traverse $\sum_{k=1}^{h} d^h$ nodes which is exponential to search depth. However, exhaustive search guarantees that all the familiar strangers of a node are found.

Random Search

The search starts from u and propagates by randomly selecting some nodes at each hop. A user-specified selectivity fraction $\sigma \in \mathbf{R}$ and $\sigma \in [0, 1]$ controls the number of contacts randomly selected at each hop. This is different than the social identity based search because of (1) no clustering of the contacts of a node, and (2) no intelligent selection of contacts in random search approach. Exhaustive search is a special case of random search where $\sigma = 1$.

Table 1: Summary of BlogCatalog and DBLP datasets.

Statistics	*BlogCatalog*	*DBLP*
Number of nodes	23,566	35,001
Number of node-node links	1,165,622	1,067,447
Link density	0.002	0.0009
Average degree of nodes	98	9
Diameter of the network	5	10
Attribute name	*Categories*	*Venues*
Size of domain of the attribute	60	3198
Average size of attribute-value set per node	1.6	28.7

Datasets

A blogger network, BlogCatalog[5] and citation network DBLP[6] is used for the evaluation of different approaches

BlogCatalog A blog in BlogCatalog is associated with various information pieces like the categories the blog is listed under, blog level tags, snippets of 5 most recent blog posts, and blog post level tags. A blogger also specifies his social network of other bloggers. A blogger's interests could be gauged by the categories he publishes his blogs in. There are in total 60 categories in BlogCatalog. Each blogger could list his blog under more than one categories. On average each blogger lists their blog under 1.6 categories. All the categories his blog has been published are agglomerated to construct his profile vector. This profile vector forms the attribute-value set for this blogger. However, in case where the category information is unavailable, we can use various existing author-topic model extraction approaches (Rosen-Zvi et al. 2004) to extract topics of the author from the text in blog posts, tags, and comments. Note that the blogger's social network vector is extremely sparse as also depicted by the average degree of nodes and link density in Table 1.

DBLP dataset presents information on computer science publications. We construct social network of authors using the co-author relation. Two authors are connected through an edge if they have collaborated on at least one paper. So all the co-authors of an author constitute his social network. Each author publishes his work in the choice of his venue, which also tells us about his interests. Based on the venue information of the publications we construct the attribute-value set of each author. We use a part of DBLP dataset which is the largest connected component of the graph generated using the co-author relation. The average degree of the author social network is 9. This shows that on average an author collaborates with 9 authors, much smaller than the BlogCatalog dataset, due to which the diameter is twice as large as the BlogCatalog as summarized in Table 1.

Dataset Characteristics

For BlogCatalog and DBLP, we investigate characteristics like power-law degree distribution and small-world assump-

[5] http://www.blogcatalog.com
[6] http://kdl.cs.umass.edu/data/dblp/dblp-info.html

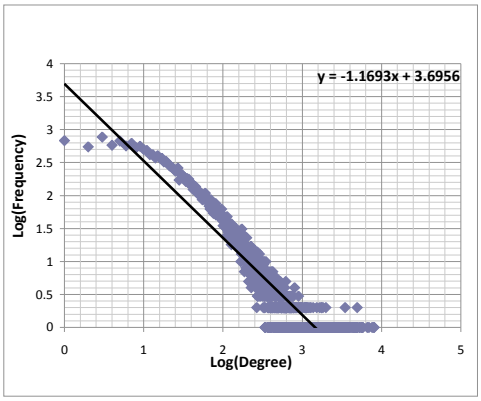

Figure 2: Degree distribution for BlogCatalog.

tion which are necessary for searchability in the network with local information (Watts, Dodds, and Newman 2002).

Degree Distribution We study the degree distribution of the nodes in BlogCatalog and DBLP dataset. We display the log-log graph of this distribution with $\log(degree)$ on the x-axis and $\log(frequency)$ on the y-axis, for BlogCatalog in Figure 2. We omit the degree distribution plot for DBLP due to space constraints. We observe that both BlogCatalog and DBLP dataset follow power law distribution $P(x) \sim x^{-k}$ with scaling exponent k of 1.1693 and 2.7896, respectively.

Small-World Assumption Networks conforming to small world assumption are characterized by short average path lengths and high clustering coefficient (Watts and Strogatz 1998). The distance between any two nodes in the network is defined as the number of edges along the shortest path connecting them. Average path length of a network is defined as follows (Watts and Strogatz 1998):

$$l_G = \frac{1}{n \times (n-1)} \times \sum_{i,j} d(v_i, v_j) \tag{1}$$

where n is the number of vertices in the graph G and $d(v_i, v_j)$ denotes the shortest path between two nodes v_i and v_j. For BlogCatalog and DBLP, we computed the average path length using the above formulae and was found to be 2.379 and 5.083, respectively.

Clustering coefficient is a common property of social networks representing circles of friends in which every member knows every other member. If a node v in graph G is connected to k_v other nodes then the clustering coefficient of node v is defined as (Watts and Strogatz 1998):

$$C_v = \frac{2E_v}{k_v(k_v - 1)} \tag{2}$$

where E_v is the actual number of edges that exist between the k_v vertices. We compute C_v for all the vertices v of the graph G and compute the average value. We compare the clustering coefficient values of the two datasets with that of random networks generated using the same set of nodes as in BlogCatalog and DBLP but the edges are rewired according to Erdös-Rényi model(Erdös and Rényi 1959). We report the results for clustering coefficient for both the datasets and their random network counterparts in Table 2, which shows that clustering coefficient values for the original datasets (Actual Networks) is much higher than their random counterparts (Random Networks). Low average path length and high clustering coefficient implies that the two datasets indeed exhibit small- world characteristics.

Table 2: Clustering coefficient results for both datasets.

	Actual Network	*Random Network*
BlogCatalog	0.51	0.001 ± 0.0002
DBLP	0.69	0.001 ± 0.0002

Experiments - Constructing Social Identity

Social identities of the nodes are not available in the online social networks, so we construct the social identities of the nodes using conventional clustering algorithm - *k-means*. BlogCatalog dataset has very rich metadata for the bloggers, including blog posts and tags. We construct the social identities of the bloggers using the metadata as mentioned in the section on Social Identity Construction. The DBLP dataset doesn't have any details about the authors besides their venues. So we cluster the contacts of an author using the venue information.

Here we present the results of social identity construction of the nodes of the blogger network from BlogCatalog dataset. To avoid the high-dimensionality and synonymy and polysemy issues we use latent semantic analysis to transform the term space vector to concept space as mentioned before in the Social Identity Construction section. Since we use k-means algorithm to construct the clusters, we need to find the optimal value of k to compute the clusters. To determine the cluster number k, we try to maximize the following ratio:

$$\frac{\frac{1}{k}\sum_{c_i}\left(\frac{2}{\|c_i\|\times(\|c_i\|-1)}\sum_{v_m\in c_i, v_n\in c_i} Cosine(v_m, v_n)\right)}{\frac{2}{k(k-1)}\sum_{c_i,c_j,i<j}\left(\frac{1}{\|c_i\|\times\|c_j\|}\sum_{v_m\in c_i}\sum_{v_n\in c_j} Cosine(v_m, v_n)\right)} \tag{3}$$
$$s.t.\ 2 \le k \le \|D\|$$

In the above formula, c_i, c_j represent two different clusters i and j. v_m, v_n are two different vectors representing two different bloggers. k varies from 2 to the number of contacts a node has, i.e. $\|D\|$. $Cosine(b_i, b_j)$ gives the cosine similarity between the two bloggers, b_i, b_j. Each blogger has two vectors: the content vector (b_i^c, b_j^c) and tag vector (b_i^t, b_j^t). We compute the cosine similarity between the two bloggers by linearly combining the cosine similarity of each of the two corresponding vectors by assigning 0.3 and 0.7 weight to content and tag vector respectively[7]. The numerator is the average similarity within the clusters, and the denominator is the average similarity between different clusters. We call them Within Similarity and Between Similarity, respectively. We plot the differential i.e., $\frac{d}{dk}\frac{WithinSimilarity}{BetweenSimilarity}$ for different values of k averaged over 100 nodes in Figure 3. We fit a polynomial trendline to help visualize the trend of the increase in the ratio of Within Similarity and Between Similarity. It is evident from Figure 3 that after a certain value of k (= 30), the increment in this ratio is small. This means that the ratio increases faster when

[7]These values of weights give the best result. Due to space constraint we do not present the results with different weight values.

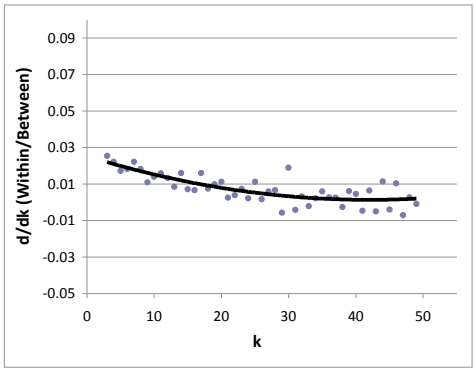

Figure 3: Differential of the ratio of Within Similarity and Between Similarity vs. *k*.

k is small, and the trend becomes flat for larger values of k. We simply set the number of clusters to 30.

To evaluate the effectiveness of k-means, we cluster the contacts by k-means and random partition, by setting the k to 30. For k-means, we randomly choose the nodes to start clustering. For random partition, the contacts are distributed into 30 clusters randomly. The average Within Similarity and Between Similarity values are computed for the clusters obtained from both k-means and random partition. Table 3 shows the average Within Similarity and Between Similarity values for k-means and random partition method over 100 runs. It is evident from Table 3 that k-means clustering gives dense or cohesive and well-separated clusters as implied by higher Within Similarity and lower Between Similarity as compared to random partition.

Experiments - Searching Familiar Strangers

In this section we compare the proposed social identity based search approach with other alternatives, viz., Steiner tree approach, exhaustive search approach and random search approach. We compare these approaches in terms of accuracy and search space complexity as explained next.

Evaluation Criteria

To compare the above-mentioned approaches we need to establish a ground truth. As mentioned in the Problem Formulation section, Steiner Tree based approach has the global view of the network $\mathcal{G}$ with V vertices, so we construct the ground truth using Steiner Tree based approach. For a given goal γ, Steiner Tree based approach extracts a subgraph $\mathcal{G}'_\gamma$ from the original graph containing nodes that share a part or whole of the γ (terminal vertices), V'_γ, as well as some nodes that do not share γ at all (Steiner vertices or optional vertices), V^{SV}_γ. This subgraph could be used to identify the familiar strangers of any node which is a part of this subgraph. Basically, the required nodes that are not directly connected to a node u in this subgraph are the familiar strangers of u or T_u and form the ground truth, denoted by V^{FS}_γ, is computed as $V'_\gamma - V^{SV}_\gamma$.

Accuracy To evaluate an approach E (where E could be one of the social identity based search approach, random search approach, and exhaustive search approach), we pick a node u from the given network such that the attribute values A_u of u and the goal γ are similar. This constraint is realized by setting $\gamma \subseteq A_u$ as also defined in the Problem Formulation section. Recall that this is the same γ that was used to generate the ground truth of familiar strangers using Steiner Tree based approach. Then we use the strategy E to generate the familiar stranger nodes for u denoted by $V^E_{u,\gamma}$. We repeat this process for all such possible nodes and aggregate the familiar strangers identified for each node, denoted by $\bigcup_{u \in V, \gamma \subseteq A_u} V^E_{u,\gamma}$. Then accuracy for approach E is computed as the intersection between the ground truth computed by using Steiner Tree based approach and the familiar strangers identified by E for the γ normalized by the total number of the familiar strangers identified by the Steiner Tree based approach as the ground truth. Mathematically, we can represent accuracy of an approach E with respect to a goal γ as,

$$Acc^E_\gamma = \frac{|V^{FS}_\gamma \bigcap (\bigcup_{u \in V, \gamma \subseteq A_u} V^E_{u,\gamma})|}{|V^{FS}_\gamma|} \quad (4)$$

Search Space Complexity We define the search space complexity of an approach E as the number of hops traversed to find the set of familiar stranger nodes with respect to a goal γ ($\bigcup_{u \in V, \gamma \subseteq A_u} V^E_{u,\gamma}$). Since Steiner Tree based approach finds the set of familiar stranger nodes with respect to a goal γ by traversing minimum number of edges. We exploit this property to establish the lower bound on the search space complexity for various approaches.

Table 3: Within Similarity and Between Similarity by different clustering methods

	K-means	*Random*
Within Similarity	0.71	0.52
Between Similarity	0.51	0.52

Table 4: Comparison of the approaches in terms of Accuracy and Search Space Complexity for BlogCatalog dataset.

Approach (E)	*Accuracy (%)*	*Search Space Complexity (edge traversals)*
Steiner Tree	100%	$3,565 \pm 560$
Exhaustive	100%	$4,531,967 \pm 891,831$
Random	$1.0283\% \pm 0.862$	$1,823 \pm 1,833$
Social Identity	$79.2908\% \pm 9.052$	$6,032 \pm 2,117$

Results and Analysis

We test for 1000 goals (γ) which is determined by seed bloggers u. For each value of γ we generate the set of familiar stranger nodes using the approaches mentioned above. We compute the accuracy for each of the mentioned approaches as explained in the section on Accuracy and also compute the search space complexity in terms of the hops traversed as described in the section on Search Space Complexity. We average the accuracy values over all the goals, i.e., 1000 γ values. We report the average accuracy along with the search space complexity for all the approaches in Tables 4 and 5.

From the Tables 4 and 5 it can be observed that, though exhaustive approach gives 100% accuracy it bears an over-

whelming search cost to discover all the familiar stranger nodes. On the other hand social identity based search approach achieves 79.2908% accuracy for BlogCatalog and 91.3495% accuracy for DBLP dataset. However, the social identity based approach searches approximately 0.1331% of the space as searched by exhaustive search approach in BlogCatalog and 1.339% for DBLP. This shows a phenomenal reduction in the search space using social identity of the nodes while searching for familiar strangers.

We present the results for Steiner Tree based approach as a lower bound for search space complexity in Tables 4 and 5. Since Steiner Tree based approach assumes global information of the network, it can discover all the familiar stranger nodes, hence it achieves an accuracy of 100% for both BlogCatalog and DBLP dataset in minimum search steps. However, social identity based search approach, which does not have global information about the network, searches only a couple of factors more of the search space (precisely, 1.69 and 2.56 for BlogCatalog and DBLP, respectively). Since the social identity based search approach has egocentric view of the network, it cannot achieve 100% accuracy, but it still performs reasonably well as compared to Steiner tree approach for both the datasets.

Deeper analysis explains the reason for such a drastic reduction in search space complexity. We computed the average number of contacts selected at each hop for the social identity based search, which comes out to be 3% and 4% for BlogCatalog and DBLP dataset, respectively. This means that as few as 3% and 4% of nodes are selected on average at each hop that propagates the search at the next hop, respectively for BlogCatalog and DBLP datasets. This extremely small fraction of nodes selected at each hop is the reason why social identity based search approach has such a small search space complexity.

To test the effectiveness of the social identity based approach we compare it with the random search approach. Random search approach selects a percentage of nodes at each hop randomly and propagates the search to the next hop. This doesn't involve any intelligent selection of the contacts. For a fair comparison we assigned σ (the selectivity parameter for random search approach) as the selectivity for social identity based search approach, which was found to be 3% and 4% for BlogCatalog and DBLP dataset, respectively. A comparison of accuracy values between the random search approach and social identity based search approach (in Tables 4 and 5) clearly shows that intelligent selection of contacts based on social identity theory improves the accuracy phenomenally. Note that random search approach selects the contacts randomly at each hop so for each goal γ value we run the random search 1000 times and report average accuracy and search space complexity results for a particular γ. Finally for all the 1000 goal γ values we compute the average accuracy and search space complexity results.

Next we compare the various approaches at different accuracy values in terms of search complexity. This experiment is performed to observe the search space complexity behavior as we attempt to find increasingly larger number of familiar strangers. We report the results in Figure 4 for BlogCatalog dataset. Note that since the random search approach does not give reasonable accuracy ($< 10\%$) in both the datasets, we do not include it in these experiments. It is evident from the figure that exhaustive search based approach has an exponential behavior. The overwhelming search space complexity of the exhaustive search approach overshadows the search space complexity behavior for social identity and Steiner Tree approach. To observe the search space complexities of social identity and Steiner Tree based search approach we plot accuracy vs. $\log$ of search steps in Figure 5 for BlogCatalog. It shows that social identity and Steiner Tree based search approach are comparable in terms of search space complexity. However, exhaustive search approach is almost 2-3 orders of magnitude higher than both the social identity and Steiner Tree based approach for both the datasets. This shows that social identity based search is closer to Steiner Tree based search approach in terms of search space complexity although social identity based search assumes only egocentric view unlike Steiner Tree based search approach that assumes global view of the network. Similar behavior is observed for DBLP, however due to space constraints the results are not presented here.

Table 5: Comparison of the approaches in terms of Accuracy and Search Space Complexity for DBLP dataset.

Approach (E)	*Accuracy (%)*	*Search Space Complexity (edge traversals)*
Steiner Tree	100%	$4,752 \pm 907$
Exhaustive	100%	$909,543 \pm 162,651$
Random	$2.304\% \pm 0.1264$	58 ± 159
Social Identity	$91.3495\% \pm 4.4398$	$12,182 \pm 4,716$

Related Work

To the best of our knowledge no work uses the social identity theory to search for familiar strangers, so we review extant literature in identifying latent friends and clustering nodes of a social network.

Identifying Latent Friends Authors in (Schwartz and Wood 1993) use Social Network Analysis (SNA) to discover groups of individuals sharing the same connectivity properties of networks. Since this does not consider the textual information of the entities, it limits the applications of SNA. Authors in (Rosen-Zvi et al. 2004)(McCallum et al. 2005) use LDA and its variations to mine relationships between people based on the content. These approaches develop topic models on the documents submitted by the authors. Authors may produce several documents often with coauthors, making it unclear how the topics generated for these documents might be used to describe the interests of the authors. Moreover, it is challenging to learn the parameters in these approaches even though well-established approximation techniques exist. Considering the limitations of author-topic model based approaches to identify latent relations, authors in (Shen et al. 2006) train an SVM to predict the topics for bloggers from external topic taxonomies. Based on the topic similarity, further refined by the cosine similarity of actual blog content, similar bloggers can be recommended. As topic taxonomies keep evolving, it requires

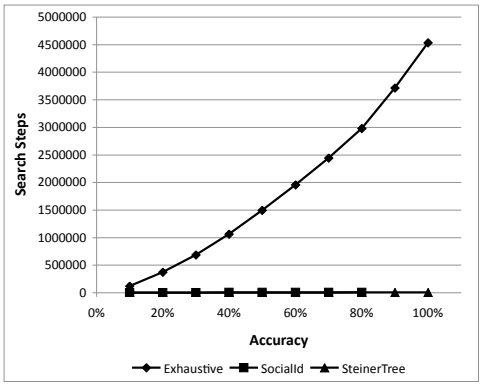

Figure 4: Accuracy vs. Search Steps for BlogCatalog.

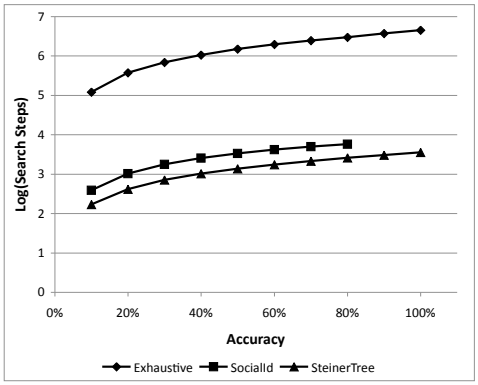

Figure 5: Accuracy vs. Log of Search Steps for BlogCatalog.

re-training the classifier that adds complexity to the solution. Moreover, detecting bloggers true interests in some of their writings could be a big challenge at times. Unlike the familiars strangers, the latent bloggers identified by (Shen et al. 2006) could possibly know each other. Other key differences are, the constraint of egocentric network view and use of social identity theory in searching familiar strangers.

Clustering in Social Networks Girvan and Newman (Girvan and Newman 2002) proposed a divisive algorithm by measuring "edge betweenness" based on the observation that the inter cluster edges have a large "edge betweenness" value if the communities are loosely interconnected. (Radicchi et al. 2004) improves the former work by considering the "edge-clustering coefficient" as the number of triangles to which a given edge belongs, divided by the number of triangles that might potentially include it, which is similar to the definition of "clustering coefficient" first introduced by (Watts and Strogatz 1998). Another measure to detect the community is modularity(Newman 2006) which estimates the fraction of in-links in a community minus the expected value of in-links in a network with the same community structure but random connections between the nodes. Unlike above methods which search for the non-overlapping communities, (Palla et al. 2005) explores overlapping communities based on the idea that a community consists of several complete subgraphs that share several nodes.

Conclusion

In this paper, we studied the familiar strangers in online social networks and identify the numerous research opportunities and business advantages of identifying and aggregating the familiar strangers. We formulate the problem and propose a social identity theory based solution with other alternatives. We also show that under certain circumstances, the problem of identifying familiar strangers can be reduced to a well-known *np*- complete Steiner tree problem and study its *2-approximation* solution to estimate the lower bound on the search space. The Steiner tree solution is also used to generate the ground truth. We performed extensive experiments on a real world blogger social network dataset, BlogCatalog and citation network dataset, DBLP to show that the proposed social identity based approach outperforms the other alternative approaches and is quite close to the Steiner tree based search approach in terms of search space complexity.

Acknowledgments

This work is in part supported by AFOSR FA95500810132 and ONR N000140810477, N000140910165 grants.

References

Bialik, C. 2007. Sorry, you may have gone over your limit of network friends. *The Wall Street Journal Online*.

Deerwester et al., S. 1990. Indexing by latent semantic analysis. *JASIS* 41(6):391–407.

Du, D., and Hu., X. 2008. Steiner tree problems in computer networks. *World scientific publishing*.

Erdös, P., and Rényi, A. 1959. On random graphs,i. *Publications Mathematicae* 6:290–297.

Girvan, M., and Newman, M. 2002. Community structure in social and biological networks. *PNAS* 99(12).

Granovetter, M. 1973. The Strength of Weak Ties. *American Journal of Sociology* 78(6):1360.

Kumar et al., R. 2004. Structure and evolution of blogspace. *Communications of the ACM* 47(12):35–39.

McCallum et al., A. 2005. Topic and role discovery in social networks. In *IJCAI*.

Milgram, S. 1972. The familiar stranger: An aspect of urban anonymity. *Division 8, Newsletter*.

Newman, M. E. J. 2006. Modularity and community structure in networks. *PNAS* 103(23):8577–8582.

Palla et al., G. 2005. Uncovering the overlapping community structure of complex networks in nature and scociety. *Nature* 435.

Radicchi et al., F. 2004. Defining and identifying communities in networks. *PNAS* 101(9).

Rosen-Zvi et al., M. 2004. The author-topic model for authors and documents. In *Proceedings of UAI*.

Schwartz, M., and Wood, D. 1993. Discovering shared interests using graph analysis. *Commun. ACM* 36(8).

Shen, D.; Sun, J.-T.; Yang, Q.; and Chen, Z. 2006. Latent friend mining from blog data. In *ICDM'06*.

Tajfel, H. 1978. *Differentiation between social groups: studies in the social psychology of intergroup relations*. Academic Press.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis*. Cambridge University Press.

Watts, D., and Strogatz, S. 1998. Collective dynamics of 'small-world' networks. *Nature* 393(6684).

Watts, D.; Dodds, P.; and Newman, M. 2002. Identity and Search in Social Networks. *Science* 296(5571):1302–1305.