Mapping Search Relevance to Social Networks

Jonathan Haynes
Department of Sociology
Stanford University
jhaynes@stanford.edu

Igor Perisic
Search, Cloud, and Data Platform Team
LinkedIn
iperisic@linkedin.com

ABSTRACT

This paper explores how information contained in the structure of the social graph can improve search result relevance. Traditional approaches to search include scoring documents for relevance based on a set of keywords or using the link structure across documents to infer quality and relevance. All these approaches have one thing in common – they attempt to optimally match keywords to documents with little or no information about the searcher and no information about his network. This paper explores an alternative approach where 3.8M profile search queries from a large social networking site are studied in conjunction with the tie structure of a 21M member social graph. The key finding is that a quantifiable measure of social distance, when used in conjunction with standard search relevance methods, dramatically improves overall search result relevance.

Categories and Subject Descriptors

F.2.2 [**Nonnumerical Algorithms and Problems**]: Sorting and searching; G.2.2 [**Graph Theory**]: Graph algorithms; H.3.3 [**Information Search and Retrieval**]: Clustering; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Algorithms, Experimentation, Theory

Keywords

Community analysis, social search, social network analysis, search relevance

1. INTRODUCTION

Search behavior on social networking sites offers a unique opportunity for studying the connection between search relevance and social network structure. Searchers are looking for members based on their profiles – either explicitly by name or with some combination of company, title, industry, interests, affiliation, or profile keywords. Unlike web search engines, social networking sites have additional context about a searcher. In other words, instead of implicit contextual, geographic or demographic data, members typically provide personal or resume style information. Social networking sites also have the social graph of member connections. This makes it possible to take member-pair attributes into account as well as their relative positions within the social graph when evaluating likely relevance to the searcher across many potential matching results.

The 3rd SNA-KDD Workshop '09 (SNA-KDD'09), June 28, 2009, Paris, France.

1.1 Background

Your social network is a unique reflection of you. Many studies of large social networks show us that social networks tend to exhibit high levels of local clustering (you tend to know your friend's friends), yet also exhibit short average path lengths. [19]

People tend to share at least one dimension of social life in common with each of their contacts – hence the context for a personal or professional relationship. [12, 2] People also tend to share similar sets of cultural and consumer preferences with their close social contacts. [6, 11]

This is due to:

- Homophily - People seek out others with similar interests
- Diffusion - People within a group are exposed to similar ideas
- Social Identity - Cultural preferences or shared experiences signal group memberships

Social network structure is not arbitrary, and in fact, the structure itself represents useful information. By design, social networking sites often return result sets consisting mostly of profiles three degrees or less away from the searcher or within similar affiliation networks, implicitly recognizing that social distance is related to relevance. As one would expect, most results are at least three degrees away. [Figure 1] In these cases, the keywords are often insufficient to differentiate which of the potential result profiles may be the most relevant.

However, estimating social distance between individuals is more than just geodesic path length or counting the number of geodesic paths, as random graphs with even a small proportion of distant ties will exhibit short average path lengths and many short paths.

1.2 Motivation

There is a good reason to believe the social graph structure should be particularly useful for people search, because the individuals we search for implicitly have some type of personal relevance to us. Many of these individuals are therefore more likely to be closer to us in social space than randomly selected members with similar characteristics.

For example, say two individuals search for a common name with the same name keywords. In such a case, they may well be searching for two different people. Each individual is likely searching for the one with closer proximity to himself in social space, because that person is more likely to be relevant due to a combination of similar geography, industry, education, and more importantly, abstract characteristics, like interests, culture, or shared identity. This is particularly true for more ambiguous searches such as a first name and an attribute, like *Elizabeth* and *P&G*, or *Matt* and *computer science*. Such searches are common. The challenge is determining a method for estimating social distance which produces useful results.

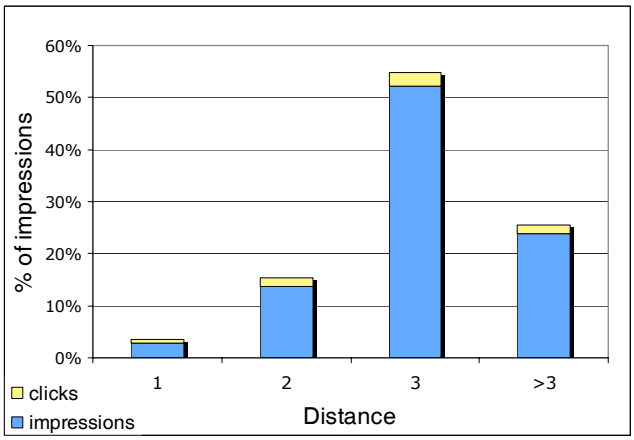


Figure 1: Frequency of impressions for search result profiles using keyword relevance by geodesic distance from the searcher

1.3 Theory

Proposition 1: Search relevance, particularly for people, is affected by social characteristics of the searcher.

This study conceptualizes *social distance* as a measure of relative proximity within social structure. Social structure is usually thought of as enduring patterns of behavior within society, as recursively implicated in the reproduction of social systems. [4] While the term social structure implies much more than just social network structure, the two concepts are closely related and, for the purposes of this study, are used interchangeably. Social distance also relates to identity, which is usually discussed in terms of overlapping group affiliations within society. [16]

For simplicity, *social distance* is operationalized here as mutually exclusive shared group memberships. All members in the social graph are clustered based solely on tie structure and each member is assigned to one subgroup. The scope of this paper does not consider other types of assignment such as simultaneous membership in multiple subgroups. Therefore, all pairs of members are either in the same subgroup (socially similar) or different subgroups (socially distant). This approach captures abstract dimensions of similarity between members that would be difficult or impossible to capture using member attributes due to limited profile information. (*see also* Kleinberg 2001 [7])

Here is a concrete example: Sarah, a software engineer with Microsoft in Dublin Ireland, and Ruth, a medical researcher at UCSF in San Francisco California both search for "John Smith". Sarah and Ruth have no friends or acquaintances in common. "John Smith" is a common name. Given this, odds are Sarah and Ruth are not looking for the same person. In such a case, keyword relevance alone is insufficient and perhaps even misleading since relevance is not absolute. Note also, geodesic distance is also insufficient since short average path lengths are typical for interpersonal social networks. [Figure 2]

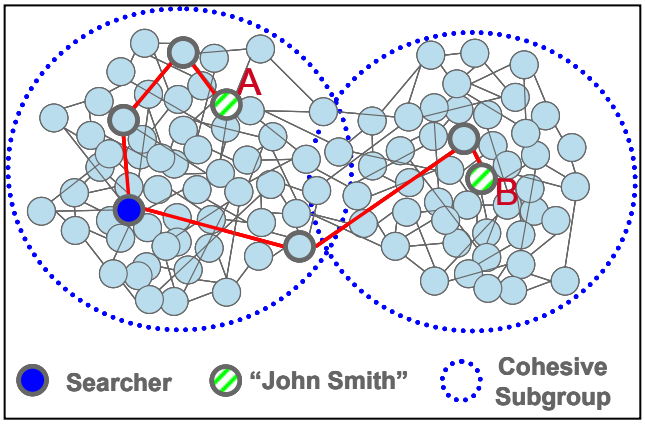


Figure 2: An illustrative example where there are two equally good keyword matches for a name search – *A* and *B* – and both are three degrees from the searcher. However, *A* is the better match as *A* is in the same subgroup as the searcher.

Therefore, there are sound theoretical reasons to believe that social network structure can be used to improve the ordering of search query results. This question has implications beyond search relevance. If the structure of the social graph can improve search relevance, it can also be used to improve site content relevance more generally.

2. RELEVANT WORK

2.1 Empirical studies of large-scale networks

Many excellent empirical studies of large-scale social networks have appeared over the past couple years. Some notable studies include the following:

Leskovec and Horvitz (2007) have conducted the largest social network study to date. [10] They study anonymized data for one month of high-level communication activities across the whole Microsoft Messenger instant-messaging system. In particular, they examine characteristics and patterns of collective dynamics of a system with 30 billion conversations (no message content) and 240 million people, which produced a graph of 180M nodes and 1.3B edges. They confirmed Milgram's six degrees of separation for IM, noting the average path length among users was 6.6 degrees.

Leskovec, Backstrom, Kumar and Tomkins (2008) have studied network evolution across four large online social networks:

Flickr, Delicious, Yahoo! Answers, and LinkedIn. [9] They propose the use of model likelihood of individual edges to evaluate and compare various network evolution models.

Onnela, Saramäki, *et al*. (2007) examine the communication patterns of millions of mobile phone users. [15] They conduct an in-depth analysis of the local and global structure of a society-wide communication network. They find that network structure is robust to the removal of strong ties but falls apart after a phase transition if the weak ties are removed. This is consistent with Granovetter's strength of weak ties thesis that weak ties are important for the global stability of a social network. [5]

The present paper contributes to this existing literature as it is the first to study search behavior in conjunction with a social graph composed of tens of millions of members.

2.2 Search in social networks

Search has been a topic of interest to social network scholars for many years. In 1967, the social psychologist Stanley Milgram tested the hypothesis that any one person in the world could be reached through a network of acquaintances in only a few steps. For the purposes of his study, an acquaintance is anyone known to someone on at least a first name basis. This question became known as the small world problem: "Given two individuals selected randomly from the population, what is the probability that the minimum number of intermediaries required to link them is 0, 1, 2,…*k*?" [17] When Milgram asked people how many steps it would take on average to get a letter from one arbitrary person in one place to an arbitrary person somewhere else, typical estimates were in the hundreds. The result was six.

The initial paradox of social networks is that on the one hand they are highly clustered, but on the other we can still manage to reach almost anyone in only a few steps. Implicit to Milgram's study is a question of search. Kleinberg has followed up on this by asking the question: given that short chains exist, how is it arbitrary pairs of strangers are able to find short chains of acquaintances that link them together? [8] He concluded that social networks contain cues fundamental for finding paths through a network. In other words, the structure of socially distance connections forms a type of "gradient" that helps individuals guide a message efficiently toward a target.

This paper addresses a different type of search behavior than search in small worlds, as the goal here is to improve the ordering of search results for profile pages based on network structure as opposed to identifying short paths to pass a message through a network.

2.3 Clustering algorithms for large graphs

The simplest form of network clustering is based on connectivity. However, simply identifying components is insufficient for interpersonal social networks, as a single giant component usually dominates large graphs. Methods based on graph-theoretic features or iteratively assigning nodes to groups until an optimum index of clustering is found are often not practical due to computational complexity constraints.

Two methods that work well for graphs larger than 10M vertices are 1) Moody's recursive neighborhood mean (RNM) algorithm (2001) and 2) Heuristic methods based on modularity optimization. Modularity optimization approaches tend to be based on models initially proposed by Newman (2004) and Clauset, Newman, and Moore (2004). Notably, Blondel, Guillaume, *et al*. (2008) and Wakita and Tsurumi (2007) demonstrate highly scalable variants. [13, 14, 3, 1, 18] While all of these are useful algorithms, Moody's approach is used here since it is fast, efficient, and based on a peer influence model which works well for identifying social groups.

3. METHODOLOGY

Clustering, like all unsupervised learning, is a descriptive method for finding groups within a dataset where the true number of groups is initially unknown. This research builds on prior work by using search behavior – both search queries and search result clicks – to determine optimal parameters for a clustering algorithm with the objective of producing socially meaningful groups of actors. In other words, the groups need to be as large as possible but small enough to contain members who are more socially similar to others in the same group as opposed to across groups. Social similarity is often difficult to measure, but in this context is simply defined as relevance – *i.e.* higher click through rates – based on a member's search behavior.

3.1 Research Objectives

1) Identify cohesive subgroups within the social graph
2) Test if subgroup membership is correlated with search result clicks

3.2 Data

The data set is from LinkedIn, a large professional social networking site. There are three separate components to the data set: a) 3,835,364 search queries (which had results) for one week in March 2008 with 38,121,383 result impressions, b) the social graph for 20,856,879 members who had at least one connection, and c) self-reported industry of employment and approximate geographic location for members. Note, this study focused exclusively on aggregate patterns of search behavior on the site. The dataset was stripped of all member identifiers and personal information before aggregate patterns were analyzed.

Forty-three percent of search queries resulted in at least one click (1,662,258). There were 1,315,469 sessions with three searches on average per session and 898,598 sessions with at least one click (68.3%). Approximately three percent of active members on the site used search during this one-week period. [Table 1]

Keyword search counts are much higher than other types of searches since these are cases where users enter any search terms into the search box. All other search characteristics, *e.g.* company, last name, etc., are features of advanced search, so users must navigate to advanced search to conduct a more structured search.

Table 1: Types of Search Queries

Search Characteristics	Counts	Frequency
Keyword(s) (incl. first name)	1,586,100	41.4%
Keyword(s) (no first name)	1,221,806	31.9%
Company	455,233	11.9%
Last Name	268,921	7.0%
Company + Title	112,848	2.9%
Title	47,622	1.2%
Keyword(s) + Company	46,119	1.2%
Other Combinations	96,715	2.5%
	3,835,364	100%

Search query sample collected for one week in March 2008

3.3 Clustering approach

We use a modified version of Moody's recursive neighborhood mean (RNM) algorithm with *k*-means to cluster based on the tie structure of the social graph. The time complexity of our modified RNM algorithm is $O(n + m)$, which makes it particularly well suited for networks with tens of millions of vertices and hundreds of millions of edges. The three required parameters are the number of position vectors (P), number of iterations (T), and number of clusters (K). We ran multiple starts with different combinations of parameters. Ultimately, we used $P = 10$, $T = 8$, and $K = 100$, as this produced stable subgroup assignments and completed in a few days on a standard workstation.

See Moody 2001:266 for a description of Moody's original RNM algorithm. [13] We modify it as follows[1]:

1. Assign each vertex in the network a uniform random value between 0 and 1 on each of p variables, $\boldsymbol{Y}$.
2. Do steps three and four t times.
3. Reset each vertex's value(s) for $\boldsymbol{Y}$ to the mean of their adjacent neighbors:

$$Y_{ipt} = \frac{\sum_{j \in L} Y_{jp(t-1)}}{|L|} \qquad (1)$$

   Where i indexes vertices, p indexes dimensions, t indexes the iteration number, and L is the set of j neighbors adjacent to i in the graph. The number of operations for each iteration is nd, where n is the number of vertices and d is the average degree in the network.

4. After each iteration, renormalize the range for each of p variables back to [0,1]:

$$\tilde{Y}_{ipt} = \frac{Y_{ipt} - \min_i(Y_{ipt})}{\max_i(Y_{ipt}) - \min_i(Y_{ipt})} \qquad (2)$$

   This step introduces differential mixing times, so that more highly connected subgraphs quickly converge to meta-stable equilibria, while sparse connections across these subgraphs lead to a slower time scale over which these meta-stable values change.

5. Run *k*-means on the p-dimensional space.

We use *k*-means for the final clustering step with RNM since it is fast and effective. However, do keep in mind that *k*-means can perform poorly for different size clusters, clusters with different densities, non-globular clusters, the wrong *k,* or outliers.

Finally, we did not use *k*-cores since the measure rarely captures sociologically meaningful groups. While this is also a scalable measure with time complexity $O(n + m)$, at higher values of k, peripheral group members are lost while highly connected members of groups may have sufficient ties across groups to merge adjacent clusters. [13]

4. RESULTS

4.1 Subgroups are geographically diverse

One might think that geographic proximity is sufficient for predicting relevance, because shared location is the most common way we interact with other people. But *Figure 3* shows that geography is an insufficient predictor. In this map, the social graph is divided into ten clusters using modified-RNM, where six clusters are of sufficient size to plot on a global map. Again, these clusters are essentially subgroups where the density of ties within a cluster tends to be greater than across clusters. You'll notice that while geography is a factor, there are many overlapping clusters.

4.2 Searchers are substantially more likely to click results in their own subgroups

In *Figure 4*, the social graph is divided into one hundred subgroups. Each column contains descriptive statistics for the ten largest distinct clusters. The lines are click through rates and the numbers at the bottom provide basic information about each cluster. The conditional probability of clicking a search result in the same subgroup is consistently double that of clicking a result in a different subgroup.

[1] Note: We removed all vertices with degree = 1 from the graph before beginning. We also removed all open networkers and recruiters since these members have many connections that do not reflect real-world relationships. We added both of these groups back at the end and assigned each to the modal subgroup membership of their network neighbors.

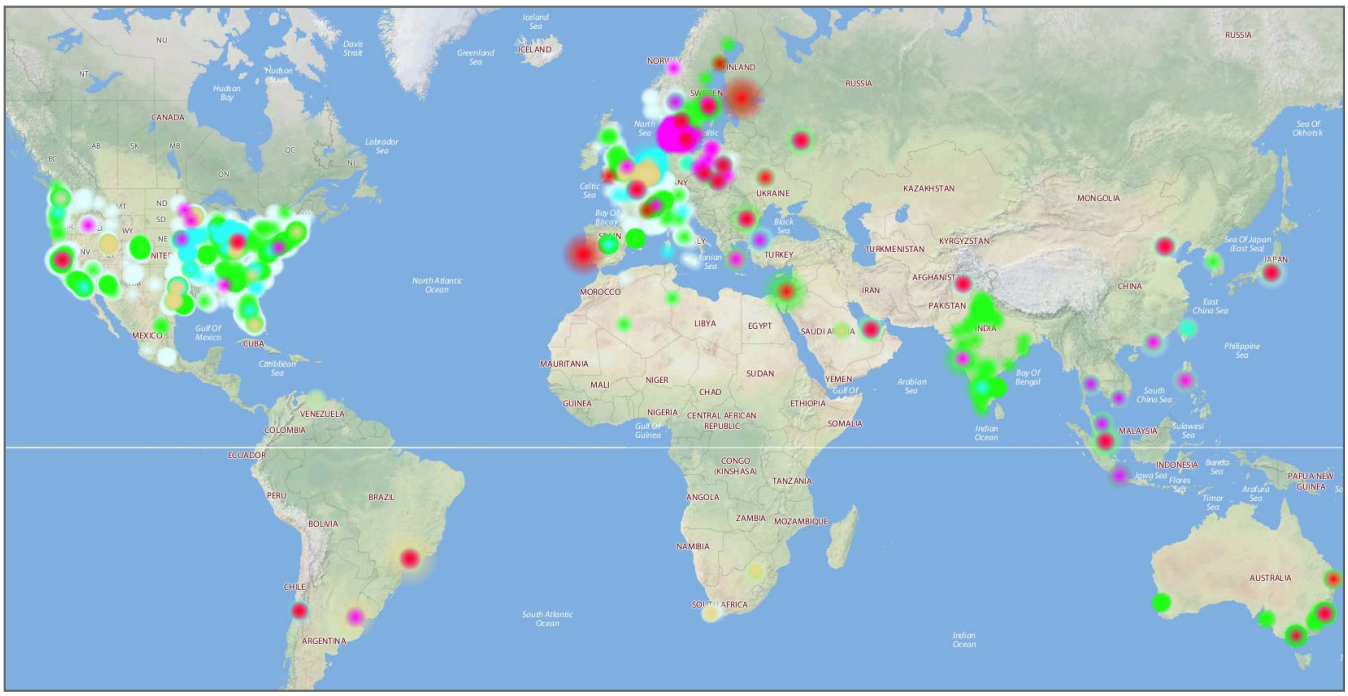


Figure 3: Geographic locations of the six largest distinct subgroups with geo-location data

n = 20,856,879 members with degree ≥ 1. This map depicts the six largest distinct modified-RNM clusters of site members based on the link structure of the social graph (K = 10, which results in six large clusters). Each color is a different subgroup. Larger circles represent more members at a given location. Geography is correlated with subgroup membership, but location alone is insufficient for predicting subgroup membership.

Notes: This map is meant as a very simple illustration. If a member's country does not have lat/long geo-location data, the largest city in the country is used. There also must be more than fifty users at a specific lat/long to appear on this map. These clusters are the best fit for K = 10, however, ten were selected to produce a clear illustration and are not in fact the best fitting number of subgroups for improving search result relevance.

This analysis is informative but insufficient, as ultimately we need to eliminate the most likely alternative hypotheses for these differences in CTRs. For example, we know that many members use search for site navigation, and since your contacts are usually in the same subgroup as you, this contributes to the lift. So, the question becomes: how much of this gain is attributable solely to group homophily as opposed to other factors?

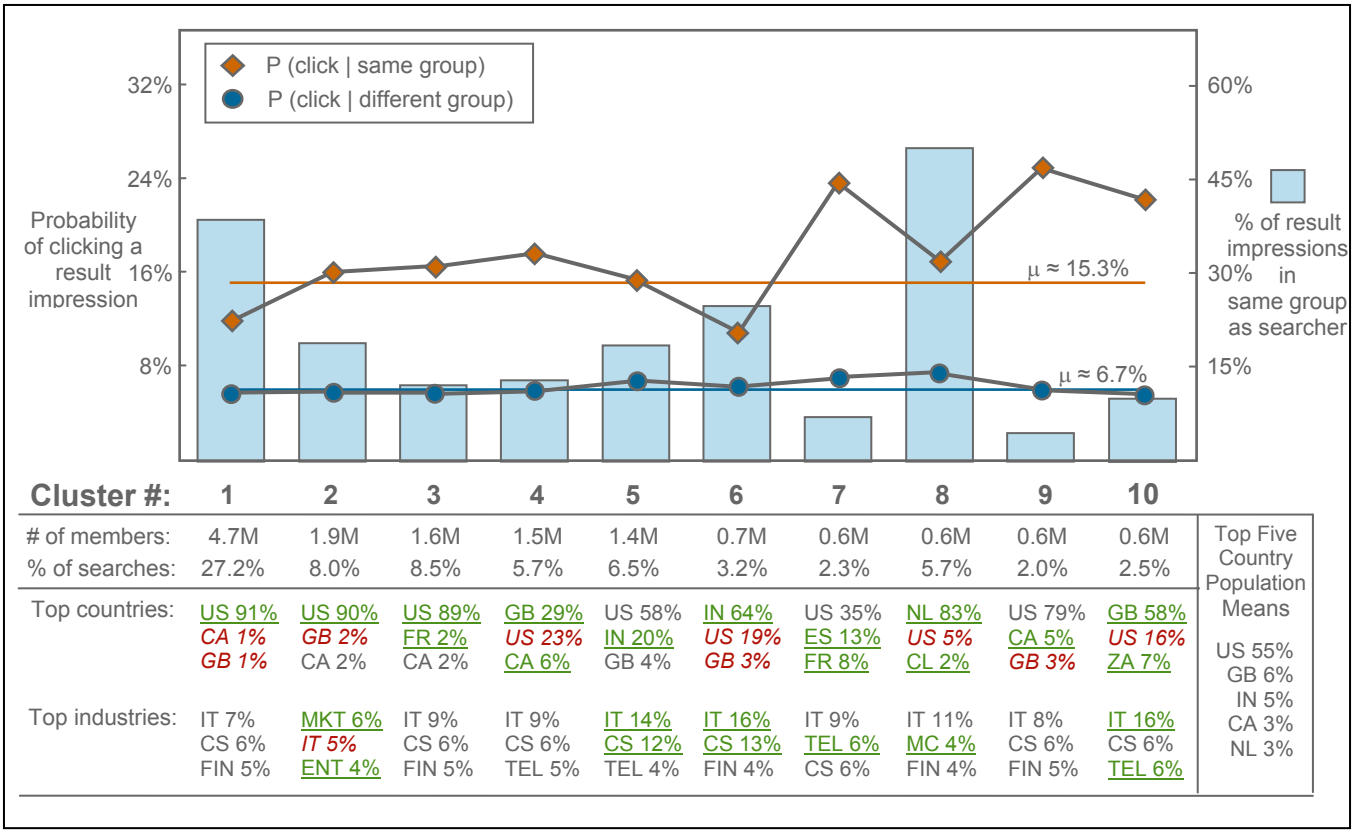
The logistic regression in *Table 2* takes multiple factors into consideration at once: position in search results (result rank), membership in the same modified-RNM subgroup, geodesic distance from searcher, and whether or not search terms include a name (type of search). The reference group for *distance from searcher* is Degree = 3. The reference group for *type of search* is a search query with no first name in the keywords field and no entry in the 'name keywords' advanced search field.

Taking into account result rank, geodesic distance from the searcher, and the type of search, searchers are 54% more likely to click results in their own subgroup. However, using just keyword relevance with no network data, only 31% of search result impressions are from the same subgroup as the searcher. Due to the large size of the sample, all coefficients are statistically significant; therefore standard errors and z-values are not included in the table.

Table 2: Logistic Regression for Clicking a Search Result Impression

Variables	Coef.	Odds Ratio
Result Rank	-0.005	0.995
Same Subgroup	0.433	1.542
Distance from Searcher		
Degree = 0	-13.09	0.000
Degree = 1	1.295	3.651
Degree = 2	0.846	2.330
Degree > 3	-0.107	0.898
Type of Search		
Name is in 1st Degree	0.612	1.844
Name is not in 1st Degree	0.628	1.874
Intercept	-3.014	

n = 10,000,000 randomly selected search result impressions; LR $\chi^2(8)$ = 253977, Prob > χ^2 = 0.0000



Cluster #:	1	2	3	4	5	6	7	8	9	10	Top Five Country Population Means
# of members:	4.7M	1.9M	1.6M	1.5M	1.4M	0.7M	0.6M	0.6M	0.6M	0.6M	
% of searches:	27.2%	8.0%	8.5%	5.7%	6.5%	3.2%	2.3%	5.7%	2.0%	2.5%	
Top countries:	US 91% *CA 1%* *GB 1%*	US 90% *GB 2%* CA 2%	US 89% FR 2% CA 2%	GB 29% *US 23%* CA 6%	US 58% IN 20% GB 4%	IN 64% *US 19%* *GB 3%*	US 35% ES 13% FR 8%	NL 83% *US 5%* CL 2%	US 79% CA 5% *GB 3%*	GB 58% *US 16%* ZA 7%	US 55% GB 6% IN 5% CA 3% NL 3%
Top industries:	IT 7% CS 6% FIN 5%	MKT 6% *IT 5%* ENT 4%	IT 9% CS 6% FIN 5%	IT 9% CS 6% TEL 5%	IT 14% CS 12% TEL 4%	IT 16% CS 13% FIN 4%	IT 9% TEL 6% CS 6%	IT 11% MC 4% FIN 4%	IT 8% CS 6% FIN 5%	IT 16% CS 6% TEL 6%	

Figure 4: Search result click-through-rates for the ten largest subgroups ($K = 100$)

n = 20,856,879 members with degree ≥ 1; Notes: These CTRs are for the subset of searches for specific people. *i.e.* name advanced search or a first name anywhere in the keywords field. Underlined figures are > 50% above a population mean. *Italicized* figures are > 50% below a population mean. These show what countries and industries are disproportionately over or under represented in a given cluster.

5. DISCUSSION

User behavior exists within a social context. A quantifiable measure of social distance, when used in conjunction with standard search relevance methods, improves search result relevance. Even after adjusting for result rank, geodesic distance from the searcher, and the type of search, users are 54% more likely to select a search result from their own subgroup.

Four areas for future research include: 1) Further work on algorithms extracting community structure from networks. While many good algorithms exist, it would be helpful to have more clustering algorithms that run in nearly linear time for interpersonal social network topologies, *i.e.* where $m > n$, as different group characteristics are desirable in different circumstances. 2) While modified-RNM and heuristic methods for modularity optimization are in some ways similar, a next step is testing relative performance with respect to search applications and identifying optimal parameters. 3) We used a simplistic measure of social distance - mutually exclusive shared group memberships – to demonstrate the potential of this approach. A hierarchical or multi-dimensional group measure should perform even better. 4) This research focused on profile searches on a social networking site. The next logical direction is identifying how to improve the ordering of search results using the social graph for other types of online search.

6. ACKNOWLEDGMENTS

Our thanks to ACM SNA-KDD reviewers for inclusion in the workshop. Thanks also to Jonathan Goldman, Jay Kreps, Mark Granovetter, Xueguang Zhou, Chandler Johnson, Jim Merino, and members of the Economic Sociology and Organizations Workshop at Stanford University for valuable feedback on earlier drafts.

7. REFERENCES

[1] Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics* (2008) P10008.

[2] Butts, Carter T. "Predictability of Large-scale Spatially Embedded Networks" (October 1, 2002). *Institute for Mathematical Behavioral Sciences*. Paper 1.

[3] Clauset, Aaron, M.E.J. Newman, and Cristopher Moore. "Finding community structure in very large networks." *Physical Review E*, 70:066111, 2004.

[4] Giddens, Anthony. 1976. New Rules of Sociological Method: A Positive Critique of Interpretative Sociologies. London: Hutchinson.

[5] Granovetter, Mark. "The Strength of Weak Ties." *American Journal of Sociology*, 78:6, 1973: 1360-1380.

[6] Hill, Shawndra, Provost, F. and Volinsky, C. "Network-based Marketing: Identifying Likely Adopters via Consumer Networks." *Statistical Science* 22(2): 256-276 (2006).

[7] Kleinberg, Jon. "Small-World Phenomena and the Dynamics of Information." *Advances in Neural Information Processing Systems* (NIPS) 14: 2001.

[8] Kleinberg, Jon. "The small-world phenomenon: An algorithmic perspective." *Proceedings of the 32nd ACM Symposium on Theory of Computing*. 2000.

[9] Leskovec, Jure, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. "Microscopic Evolution of Social Networks." *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM KDD)*, 2008.

[10] Leskovec, Jure and Eric Horvitz. "Worldwide Buzz: Planetary-Scale Views on an Instant-Messaging Network." *Microsoft Research, Technical Publication* MSR-TR-2006-186, June 2007.

[11] Mark, Noah. "Culture and Competition: Homophily and Distancing Explanations for Cultural Niches." *American Sociological Review*, Vol. 68, No. 3 (Jun. 2003), pp. 319-345.

[12] Mayhew, Bruce H. (1984). "Chance and necessity in sociological theory." *Journal of Mathematical Sociology*, 9: 305-339.

[13] Moody, James. "Peer influence groups: identifying dense clusters in large networks." *Social Networks*. Volume 23, Issue 4, October 2001, 261-283.

[14] Newman, M. E. J. "Fast algorithm for detecting community structure in networks." *Physical Review E,* 69:066133, 2004.

[15] Onnela, J.-P., J. Saramäki, et. al. "Structure and tie strengths in mobile communication networks." *Applied Physical Sciences*. 10.1073, April 2007.

[16] Simmel, Georg. 1964. Conflict and the Web of Group Affiliations. The Free Press.

[17] Travers, Jeffrey and Stanley Milgram. "An Experimental Study of the Small World Problem." *Sociometry*. 1969. 32:4, 425-443.

[18] Wakita, Ken and Toshiyuki Tsurumi. "Finding Community Structure in Mega-scale Social Networks." *Proceedings of the 16th International Conference on World Wide Web*. May 2007.

[19] Watts, Duncan J. and Steven Strogatz. "Collective Dynamics of Small-World Networks." *Nature* 1998. 393-440.