How User Behavior is Related to Social Affinity

Rina Panigrahy
Microsoft Research
1065 La Avenida
Mountain View, CA 94043, USA
rina@microsoft.com

Marc Najork
Microsoft Research
1065 La Avenida
Mountain View, CA 94043, USA
najork@microsoft.com

Yinglian Xie
Microsoft Research
1065 La Avenida
Mountain View, CA 94043, USA
yxie@microsoft.com

ABSTRACT

Previous research has suggested that people who are in the same social circle exhibit similar behaviors and tastes. The rise of social networks gives us insights into the social circles of web users, and recommendation services (including search engines, advertisement engines, and collaborative filtering engines) provide a motivation to adapt recommendations to the interests of the audience. An important primitive for supporting these applications is the ability to quantify how connected two users are in a social network. The shortest-path distance between a pair of users is an obvious candidate measure. This paper introduces a new measure of "affinity" in social networks that takes into account not only the distance between two users, but also the number of edge-disjoint paths between them, i.e. the "robustness" of their connection. Our measure is based on a sketch-based approach, and affinity queries can be answered extremely efficiently (at the expense of a one-time offline sketch computation). We compare this affinity measure against the "approximate shortest-path distance", a sketch-based distance measure with similar efficiency characteristics. Our empirical study is based on a Hotmail email exchange graph combined with demographic information and Bing query history, and a Twitter mention-graph together with the text of the underlying tweets. We found that users who are close to each other – either in terms of distance or affinity – have a higher similarity in terms of demographics, queries, and tweets.

Categories and Subject Descriptors

G.2.2 [**Graph Theory**]: Graph algorithms, path and circuit problems

General Terms

Algorithm, Experimentation, Performance

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.

Keywords

Social networks, distance, affinity, influence, sketching

1. INTRODUCTION

Are the activities and preferences of people correlated to or influenced by their position in a social network? Do two users that are "nearby" in the network tend to have similar preferences, browse the same web sites, or speak the same language? If so, exactly what is the right measure of "social affinity"? While a natural choice may the shortest-path distance, this doesn't capture the number of paths between the pair of users. An alternate choice is the min-cut or the connectivity, which is also the number of edge-disjoint paths between two users. Furthermore, an affinity measure should preferably be robust – it should not change much with a small change in the edges of the network.

This paper introduces a social affinity measure that can be viewed as an intermediate measure between the shortest-path distance and the number of paths connecting a pair of nodes. Such a notion of affinity is not only useful for studying user behavior, but it has commercial applications in web search, targeted advertising, and collaborative filtering. For example, on a web search query, it may be desirable to bias the results towards documents that are authored by or preferred by nearby users (such as the Facebook "liked" documents). For this reason it is important that the distance measure should be easy to compute in an online fashion at run-time. Our affinity measure not only takes into account the lengths and the number of paths between two users, but is also efficient in the sense that it can be easily computed online (with some one-time offline precomputation).

The second contribution of this paper consists of two empirical studies of the relationship between social affinity and user characteristics and behavior, and a comparison of social affinity to approximate shortest path (ASP) distance, another efficiently computable measure. The first study is based on the email exchange graph between Hotmail users, their self-declared demographic information, and queries they issued to the Bing search engine. We found that users who are nearby in the email exchange graph tend to have similar demographic characteristics, and more interestingly, that their queries were more similar to each other than those of far-apart users. Moreover, we found that both ASP distance and social affinity between two users are correlated to their profile and query similarity. While ASP had a higher correlation for most user properties, affinity was better in one that changes over short distances. This suggests that the affinity measure may be better-suited for capturing short-distance

effects in a social graph, whereas ASP may be preferable for capturing long-distance effects. The second study is based on the graph induced by two Twitter users mutually mentioning each other in their tweets, using the textual similarities between their collected tweets as ground truth. We found that the affinity between two users is strongly correlated with the similarity between their tweets, while ASP distance in the mention-graph is less strongly correlated. We speculate that textual similarity between tweets is high in a short range of distances, and that affinity captures this better than ASP does.

There is a growing body of work on the impact of social influence on consumer preferences. Domingos and Richardson [10] proposed to model markets as social networks of consumers, and derived how to optimize profit by focusing on the most influential consumers, validating their model against a movie recommendation database; others (for example, Kempe et al. [14] and Leskovec et al. [17]) expanded on this work. Crandall et al. [8] studied the interplay between social influence and selection (individuals forming associations with like-minded people) based on a edit history of Wikipedia articles. Kossinets et al. [15] proposed a temporal notion of distance in social graphs, by quantifying how long it takes for information to propagate along a given edge. Cosley et al. [9] subsequently proposed a model of how influence propagates through a social network. The topic has also attracted much interest from disciplines other than Computer Science. For example, the medical community has investigated the correlation between social connections and public-health issues such as obesity [6], smoking [7], and alcohol consumption [20], as well as positive factors such as happiness [12].

Similarly, there is a large body of work on using sketching for estimating distances in graphs. Bourgain [3] showed how any graph can be embedded into a Euclidean space while preserving distances to a factor of $O(\log n)$; Matousek [19] obtained a tighter analysis of Bourgain's result. Thorup and Zwick [24] gave an alternate algorithm for estimating distances to within a factor of $2k-1$ by using sketches of size $\tilde{O}(n^{1/k})$. Sketching is also commonly used in comparing documents, see for example work by Broder et al. [5, 4]. Spielman and Teng [23] showed how a method known as graph sparsification can be used to estimate resistance across edges; improved in follow-on work [22, 1, 16]. Von Luxburg et al. [18] argued that resistance may not be a good measure of distance in large graphs.

2. MEASURES OF SOCIAL AFFINITY

As stated earlier, a good social affinity measure should take into account not only the length of the shortest path between two nodes (users) but also the number of paths and their lengths. Some candidate measures are:

1. Max flow (or min cut) between the pair of nodes
2. Commute time (or effective resistance) between the pair in a random walk
3. Using probabilities of reaching one node from another in a random walk

Unfortunately, all the above measures are very difficult to compute efficiently online. An alternative approach is to employ sketching. It involves a one-time precomputation that produces a compact "sketch" for each node in the graph (user in the social network). During the online phase, given a pair of users, it suffices to read the sketches stored for these users and perform some simple computation on these sketches. In the context of this paper, this computation produces the social distance between two users, or an approximation of it. Not all distance measures are sketchable. For example, the shortest-path distance cannot be sketched unless one is willing to make use of a gross approximation. On a graph with n nodes, if one is willing to obtain a $2k-1$-approximation to the shortest-path distance, then this can be achieved with a sketching algorithm that stores a sketch of size $O(n^{1/k})$ per node. In recent years, theoretical algorithms research has produced sketching algorithms for computing the commute-time between a pair of nodes [23, 16] – however, despite a low asymptotic time complexity, these algorithms are very complicated and perhaps have large constants hidden in the asymptotics. The social affinity measure introduced in this paper employs sketching to approximate the probability that the pair of nodes remain connected when a certain fraction of the edges are removed randomly. We compare this measure to the approximate shortest-path (ASP) measure [11], another sketch-based technique for estimating the shortest-path distance in a graph; ASP has been observed to have small additive error for certain real-world graphs.

2.1 Affinity based distance measure

The affinity measure introduced in this paper is essentially a quantification of the probability that a given pair of nodes remain connected when a certain fraction of the edges are removed randomly. For example, if 50% of the edges are removed randomly, then it is reasonable to conclude that pairs of nodes remaining in the same connected component have a higher affinity than the ones that got separated. If there is only one path of a certain length between a pair it is less likely to survive sampling as compared to a case when there are many paths of that length. Also, longer paths are less likely to survive than shorter paths. Thus, our measure accounts for both length and breadth of the set of paths. We formalize this intuition below.

Given a graph G, let G_p denote the graph obtained by sampling the edges of the graph with probability p (so $G_1 = G$). The affinity, parameterized by p, between a pair of nodes is defined as follows

DEFINITION 1. *$A_p(u,v)$ = probability that u and v are connected in G_p.*

One method to get a parameter-free version of affinity is to look at the mean (or expected) value when p is chosen from a distribution D such as the uniform distribution on the range $[0,1]$.

DEFINITION 2. $A(u,v) = E_{p\in D}[A_p(u,v)]$

The above definition captures the mean of A_p over a distribution. Alternatively, we could capture the median, or more generally a percentile quantified by a threshold θ. Formally,

DEFINITION 3. $A^\theta(u,v) = 1 - \min\{p : A_p(u,v) > \theta\}$

Thus $A^\theta(u,v)$ measures the fraction of edges from the graph that must be removed so that the probability of the two points being connected is no more than θ.

2.2 Relation to other measures

The following theorems show the connection between the affinity measure and other connectivity and distance measures. It has some relations to the shortest-path distance, to the minimum cut and to strong-connectivity, which is a variant of minimum cut. Assume that $A(u,v)$ is computed by drawing p uniformly from $[0,1]$.

We will use $d(u,v)$ to denote the shortest-path distance measure between u and v.

DEFINITION 4. *(Connectivity) A pair of nodes u,v is k-connected if there are k edge-disjoint paths between u and v. The connectivity $C(u,v)$ between a pair u,v is the maximum value of k for which the pair is k-connected. It is also equal to the minimum cut between the pair (by the max-flow min-cut theorem).*

DEFINITION 5. *(Strong Connectivity [2]) A pair u,v is k-strongly connected if it lies in a subset U of nodes, so that every pair of nodes in the subgraph induced by U is k-connected. The strong-connectivity $S(u,v)$ between a pair u,v is the maximum value of k for which the pair of k-strongly-connected.*

Use $\bar{A}_p(u,v)$ to denote $1-A_p(u,v)$. Similarly for $A(u,v)$ and $A^\theta(u,v)$.

THEOREM 6.

$$\Omega\left(\frac{S(u,v)}{\log n}\right) \leq \frac{1}{1-A(u,v)} \leq 1 + C(u,v)$$

$$lim_{p\to 0}\frac{\log A_p(u,v)}{\log p} = d(u,v).$$

$$lim_{p\to 1}\frac{\log(1-A_p(u,v))}{\log(1-p)} = C(u,v).$$

$$\theta\bar{A}^\theta(u,v) \leq \bar{A}(u,v) \leq \frac{1}{\theta}\bar{A}^\theta(u,v)$$

PROOF. Look at the minimum cut of size $C(u,v)$ between u and v. At sampling probability p, the probability that none of the edges of the cut are chosen resulting in a disconnection is $(1-p)^{C(u,v)}$, implying that $1-A_p(u,v) \geq (1-p)^{C(u,v)}$. Therefore if p is chosen uniformly from $[0,1]$ then $1-A(u,v) = E_p[1-A_p(u,v)] \geq E[(1-p)^{C(u,v)}] \geq \int_0^1 (1-p)^{C(u,v)}dp = \frac{1}{1+C(u,v)}$.

It is known that if you sample edges with probability $p = \Omega(\log n/S(u,v))$ then the pair (u,v) remains connected with high probability, which means that $A_p(u,v)$ is close to 1 for such p [2, 13]. This implies $1-A(u,v) \leq O(\log n/S(u,v))$.

Note that in the $lim_{p\to 0}$, the probability that all the edges on the shortest path between u and v are sampled is $p^{d(u,v)}$, and this is the least set of edges that need to be sampled and hence is a dominating factor in the probability of staying connected.

$A_p(u,v) = \sum_{i=1}^{m} N(i).p^i(1-p)^{m-1}$ where $N(i)$ is the number of edge subsets of size i in which u and v are connected. As $p \to 0$, $\frac{\log A_p(u,v)}{\log p}$ is equal to the minimum i for which $N(i)$ is non-zero which is $d(u,v)$.

A similar argument holds when $p \to 1$. In this case the smallest set of edges that disconnects u and v has size $C(u,v)$. So $(1-p)^{C(u,v)}$ is a dominating factor in the probability of disconnection.

At sampling probability $h = \bar{A}^\theta(u,v)$, $A_h(u,v) = \theta$. So at any probability $p \leq h$, $\bar{A}_p(u,v) \geq \theta$ implying that $\bar{A}(u,v) = E_p[\bar{A}_p(u,v)] \geq h\theta = \theta\bar{A}^\theta(u,v)$.

Recall that G_p denotes a subgraph of G where each edge is sampled with probability p. Note that for any integer x, G_{px} stochastically dominates the union of x independently sampled graphs G_p, because the union will contain each edge with probability at most px implying that $\bar{A}_{px}(u,v)$ is at most the probability that u and v are disconnected in all the copies, which happens with probability $(\bar{A}_p(u,v))^x$. Since $\bar{A}_p(u,v)$ is decreasing in p, we have $\bar{A}(u,v) = E_p[\bar{A}_p(u,v)] \leq \int_{x=0}^{\frac{1}{h}-1} h(\bar{A}_{hx}(u,v))dx \leq \int_{x=0}^{\frac{1}{h}-1} h(1-\theta)^x dx \leq \frac{h}{\theta} = \frac{\bar{A}^\theta(u,v)}{\theta}$ □

2.3 Affinity Sketching Algorithm

We now show that the affinity measure is easily sketchable. The algorithm computes a short sketch (a summary) per node as a precomputation offline. At runtime the sketch of two nodes is used to estimate their affinity $A(u,v)$.

The rough idea of the off-line sketch generation phase is to sample the edges of the graph at different probabilities and record the connected component a node belongs to. The probabilities could be values chosen geometrically such as $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots$ or from some other discrete set of values between 0 and 1. For each value of probability we can sample the edges of the graph and record the connected component of each node (e.g. the ID of a canonical "representative" node in the connected component). We repeat this for all probabilities, a few times for each value. The sketch $C(u)$ is simply the vector (or matrix) of component IDs for this node u over all these sampling experiments.

At runtime, to estimate $A^\theta(u,v)$ we retrieve $C(u)$ and $C(v)$. For each sampling probability find the fraction of times they were in the same component. We find the two consecutive sampling probabilities p and p' that cross over the threshold θ (one can ensure that these fractions are monotonically decreasing in the sampling probability by making the sampled sets as a telescoping sequence of sets one contained in the next one). $A^\theta(u,v)$ is concentrated between p and p'.

Now, let us give a formal definition of an affinity sketch. Let (V,E) be a social graph with vertex (user) set V and edge (relationship) set E, such that $E \subseteq V \times V$. For simplicity, we assume that the graph is undirected and that the edge set reflects this, i.e. $(u,v) \in E \equiv (v,u) \in E$.

We write $\text{Conn}(v,v',V,E)$ to denote that v and v' are connected in graph (V,E). We define a partitioning of a (sub)graph (V,E) into connected components as follows:

$\text{Part}(V,E) = \{V_1, \cdots, V_t\}$ such that

1. $V_1 \cup \cdots \cup V_t = V$
2. $\forall V_i, V_j s.t. V_i \neq V_j : V_i \cap V_j = \emptyset$
3. $\forall V_i \forall v,v' \in V_i : \text{Conn}(v,v',V,E)$
4. $\forall V_i, V_j s.t. V_i \neq V_j : \forall v \in V_i, v' \in V_j : \neg\text{Conn}(v,v',V,E)$

We define the component of a vertex v given a partitioning of a (sub)graph (V,E) as follows:

$$\text{Comp}(v,V,E) = V_i \text{ where } V_i \in \text{Part}(V,E) \wedge v \in V_i$$

The sketch construction phase of our algorithm takes three parameters: integers q and r and a sampling probability

vector $p_1, \cdots, p_r$. Given these parameters, we generate a matrix of edge subsets

$$\begin{pmatrix} E_{1,1} & \cdots & E_{1,r} \\ \vdots & \ddots & \vdots \\ E_{q,1} & \cdots & E_{q,r} \end{pmatrix}$$

such that $E \subseteq E_{i,1} \subseteq \cdots \subseteq E_{i,r}$ (i.e. each row of the matrix is a telescoping sequence of subsets) and $|E_{i,j}| = p_j|E|$ (the cardinality of $E_{i,j}$ is p_j that of the cardinality of the full edge set E). Using this definition, we define the sketch C of a vertex v to be the matrix

$$\begin{pmatrix} c_{1,1} & \cdots & c_{1,r} \\ \vdots & \ddots & \vdots \\ c_{q,1} & \cdots & c_{q,r} \end{pmatrix}$$

where $c_{i,j} = \text{Comp}(v, V, E_{i,j})$, i.e. based on the graph partitioning induced by the edge subset matrix defined above.

In order to describe how sketches are used to estimate the affinity between two vertices, we employ Iverson bracket notation: $[a = b]$ is 1 if $a = b$ and 0 otherwise, and $[a < b]$ is 1 if $a < b$ and 0 otherwise. Using this notation, we define the affinity estimates $\tilde{A}$ and $\tilde{A}^\theta$ of two vertices v and v' with sketch matrices C and C' to be as follows:

$$\tilde{A}(v, v') = \frac{1}{qr} \sum_{j=1}^{r} \sum_{i=1}^{q} [c_{i,j} = c'_{i,j}]$$

$$\tilde{A}^\theta(v, v') = \frac{1}{r} \sum_{j=1}^{r} \left[\theta < \frac{1}{q} \sum_{i=1}^{q} \left[c_{i,j} = c'_{i,j} \right] \right]$$

By setting the sampling probabilities $p_1, \cdots, p_r$ to the values $1, (1-\epsilon), (1-\epsilon)^2, ...(1-\epsilon)^r$ where $r = O(\log m/\epsilon)$ one can obtain a $1 \pm \epsilon$ approximation to the affinity between a pair of nodes.

THEOREM 7. *There is an algorithm that estimates $\bar{A}^\theta(u, v)$ between any pair of nodes (u, v), for any constant θ within a $1 \pm \epsilon$ factor with high probability. using sketches of size $O(\log m \log n/\epsilon^2)$ per node. The sketches can be computed in time $\tilde{O}(m/\epsilon^2)$.*

PROOF. Let $h = \bar{A}^\theta(u, v)$ be the sampling probability at which the probability that u and v are connected is θ. If a sampling probability p is more than $h(1+\epsilon)$ then we will argue that $A_p(u, v) \geq \theta(1 + \Omega(\epsilon))$: Note that the graph G_p stochastically dominates the union of the graph G_h and $G_{\epsilon h}$ as the sampling probability in the former is no lower than the sampling probability in the union. So $A_p(u, v)$ is at least the probability that u and v are connected in either of G_h or $G_{\epsilon h}$ which is $1 - (\bar{A}_h(u, v))(\bar{A}_{\epsilon h}(u, v))$. But again the union of $(1+\epsilon)/\epsilon$ independent copies of $G_{\epsilon h}$ stochastically dominates G_h as the sampling probability in the union is no lower. So $\bar{A}_h(u, v) \geq (\bar{A}_{\epsilon h}(u, v))^{(1+\epsilon)/\epsilon}$. This gives, $A_p(u, v) \geq 1 - (\bar{A}_h(u, v))^{1+\epsilon/(1+\epsilon)} \geq A_h(u, v)(1 + \Omega(\epsilon)) \geq \theta(1 + \Omega(\epsilon))$. Since each sampling probability is repeated $q = \Theta(\log n/\epsilon^2)$ times, by Chernoff bounds, the observed fraction of times the pair stays connected will be more than θ with high probability. Similarly we can argue that if $p \leq h(1-\epsilon)$ then the observed fraction will be less than θ. Therefore our sampling probability vector will contain two probability values that sandwich h within a $1 \pm O(\epsilon)$ factor. By appropriately adjusting the constants in the $O(.)$ notation, we get the theorem statement. □

2.4 Affinity Implementation

We built single-machine implementations of both the off-line and the on-line phase of the affinity sketch algorithm. The off-line phase consists of several steps: First, we map user IDs into a dense integer space suitable as array indices. Second, we load the edge set of the graph into an in-memory array and permute this array, either uniformly at random or biased by the weight of each edge. Third, we perform q iterations, each computing one row of the sketch matrix $C(v)$ for each vertex v. Each iteration starts with an empty edge set, and consumes the content of the permuted edge array in order, adding each new edge to a union-find forest. Whenever a sampling threshold as specified by the vector $p_1, \cdots, p_r$ is crossed, we write the connected-component ID of each node v (which corresponds to a cell $c_{i,j}$ in $C(v)$) to disk, eventually producing $q \times r$ streams of connected component IDs. Fourth, we merge these $q \times r$ streams into a single stream of sketches, each sketch being a matrix of $q \times r$ component IDs. The online phase, given a pair of user IDs, simply converts them into two integers using the same mapping as the off-line phase, and uses each integer to seek to the right position of the sketch file and to read the sketch from disk.

2.5 Approximate Shortest Path

In Section 3, we will compare the effectiveness of the affinity measure to that of the approximate shortest path measure, another sketch-based measure of distances in graphs [11]. The ASP measure is a variant of the algorithm introduced by Thorup and Zwick [24].

The idea behind the approximate shortest path (ASP) algorithm is to sample, in the off-line phase, a small number of sets of "seed" nodes in the graph. Then, for each node in the graph, find the closest seed node in each of these seed sets. The sketch for a node simply consists of the closest seeds (one per seed set), and the distance to each closest seed. Then, in the online computation, one can use the distance to these closest seeds to estimate the distance between a given pair of nodes by checking for a common seed between the two sketches. Given a pair of nodes u and v one can estimate the distance between them by looking for a common seed in their sketches. If w is a common seed in the sketch of u and v then the distance can be estimated by adding up the distances to the common seed w.

3. EXPERIMENTS

We performed two studies to evaluate the effectiveness of the Affinity measure and to compare it to ASP distance. The first study was conducted on the Hotmail email exchange graph, and used user-supplied demographic information as well as Bing query history as its ground truth; the second study was conducted on the Twitter mention-graph, and used the text of tweets as its ground truth.

3.1 The Hotmail experiment

Our first experiment is based on three data sets:

1. An anonymized data set containing a pair of user ID hashes for any two email users (at least one of which is a Hotmail user) that had a mutual email exchange (that is, both users sent email to one another). This data set induces a graph with 312,548,443 nodes and 574,434,516 undirected edges, implying an average de-

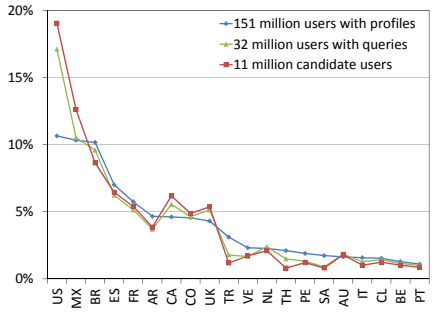

Figure 1: Country distribution of Hotmail users

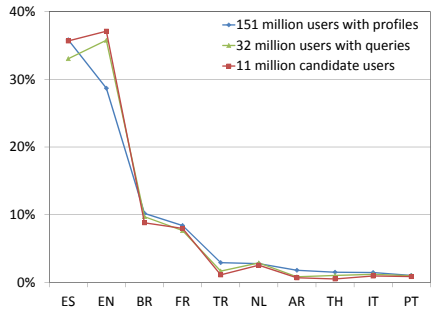

Figure 2: Language distribution of Hotmail users

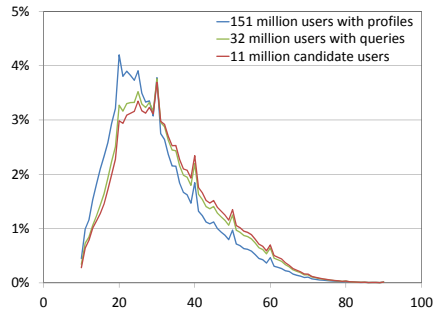

Figure 3: Age distribution of Hotmail users

gree of about 3.676. The graph is highly-connected, with 309,065,428 of the nodes (close to 99%) being part of a giant connected component. The data was collected between October 2007 and April 2010, and represents a sample of the email traffic (condensed to a hash of the sender's and receiver's email address) going through Hotmail during that timeframe. Assuming that user u sent a messages to user v and v sent b messages to u, the edge (u, v) in the graph has a weight of $\min\{u, v\}$ associated with it.

2. A data set containing basic, self-reported demographic information (including country of residence, primary language, and age) for a subset of 151,265,301 Hotmail users in the first dataset.

3. A data set containing the Bing query history for a subset of 32,212,592 users in the second data set.

We furthermore restricted this set of 32,212,592 users covered by all three data sets (that is, for whom we have email exchange history, demographic information, and Bing query history) to contain only users who had issued at least five queries so as to be able to make meaningful comparisons between two users' query histories, and we restrict who had a minimum of two email partners in order to compensate for the vagaries of email traffic sampling. This left us with a set of 10,974,103 "candidate" users.

Figure 1 shows the distribution of the top twenty countries of residence of Hotmail users (the top twenty countries being those with at least 1% of users in the overall data set). The horizontal axis shows the country, ordered by decreasing frequency in the overall data set; the vertical axis shows the percentage of users from a given country. The blue curve plots the distribution of the 151 million users for whom we have demographic information; the green curve shows the distribution of the 32 million users for which we also have queries; and the red curve shows the distribution of the 11 million "candidate" users from which we sampled pairs. While the largest portion of Hotmail users resides in the United States, a surprisingly large portion resides in Latin America (Mexico, Brazil, Argentina, Columbia, Venezuela, Peru, and Chile). It is also worth noting that users from the US are overrepresented in the users-with-queries set relative to their representation in the overall data set. This is due the fact that Bing is particularly popular in the US. Users from the US and Mexico yet more overrepresented in the candidate set, suggesting that they either issue more queries or maintain a larger set of email partners.

Figure 2 shows the distribution of the top ten primary languages of Hotmail users (the top ten languages being those with at least 1% of users in the overall data set), using the same encoding as the previous figure. The largest portion of users in the overall set are Spanish-speakers (consistent with the popularity of Hotmail in Latin America and Spain), while the users-with-queries and candidate sets are dominated by English-speakers. We attribute this to the overrepresentation of US users in the candidate set.

Figure 3 shows the distribution of (self-declared) ages of Hotmail users, using the same encoding as the previous two figures. The graph shows only ages from 10 to 90 years, which accounts for 98.4% of the overall user base. The overall user base is dominated by people in their early twenties, users in the users-with-queries and candidate sets are slightly older. Notice the peaks at ages that are a multiple of ten, and minor peaks at multiples of five, which suggests that users round their ages before reporting.

We used the implementation described in section 2.4 to compute sketches for each node in the email exchange graph, using a machine with a dual-core AMD Opteron 285 processor clocked at 2.6 GHz, 16 GB of RAM, and a RAID-5 array composed of eight 1 TB, 7200 RPM Seagate Barracuda SATA drives. We set q to 10 and r to 100, yielding sketches of 4000 bytes per node. We computed three sets of sketches: one where edges were permuted uniformly at random, one where edges were permuted biased by their weight (the count of email exchanges), and one where edges were permuted biased by the log of their weight. The computation took 14 days. Most of the time was spent on disk I/O; indeed, the final merge phase (which merged a thousand files into a single 1.13 TB file) took 11 days. We also ran a variant of this program that performs the same computation, but records statistics instead of writing out sketches. Running this variant over the same input data took slightly less than 22 hours.

Instead of choosing a geometric progression for the sampling probability vector $p_1, \cdots, p_r$ as suggested in in Theorem 7, we calibrated it such that the affinity value between a random pair of nodes is uniformly distributed in the range $[0, 1]$.

We randomly sampled 5 million pairs of users from the set of 10,974,103 candidate users, and computed the ASP distances and affinities between each pair. Using the hardware described above and the 1.13 TB sketch file produced by the off-line phase, computing the affinity value for each pair of Hotmail user IDs takes about 17 milliseconds; this time is dominated by the cost of performing two disk seeks.

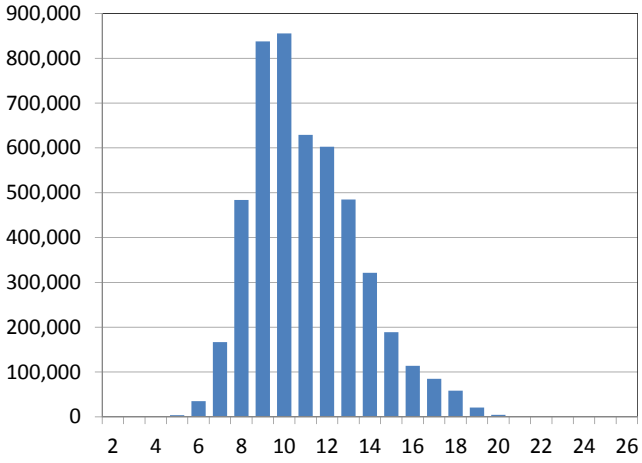

Figure 4: Distribution of ASP distances for the sampled 5 million pairs of users

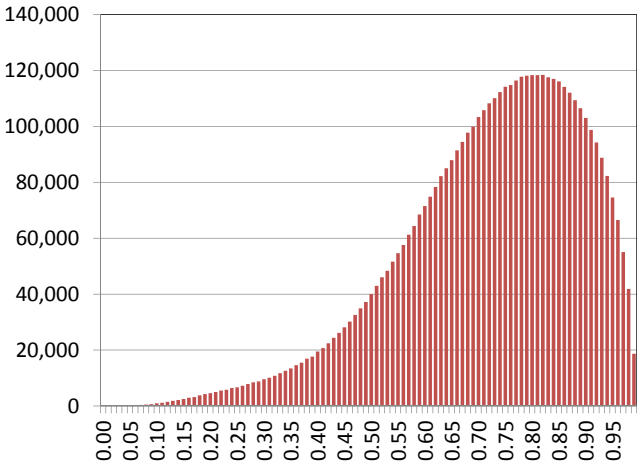

Figure 5: Distribution of affinities for the sampled 5 million pairs of users

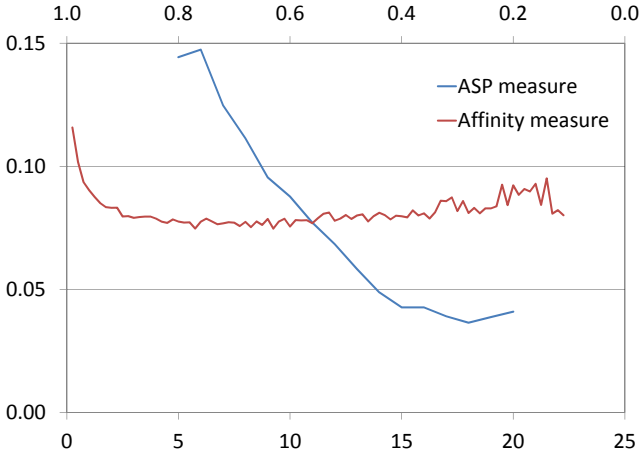

Figure 6: Fraction of pairs of users in same country as a function of ASP or affinity measure

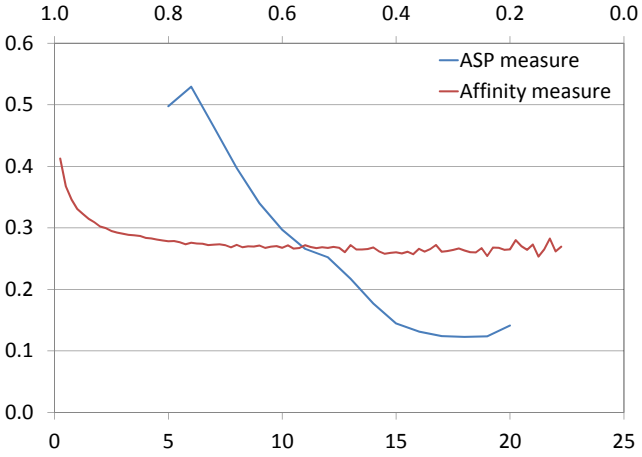

Figure 7: Fraction of pairs of users speaking same language as a function of ASP or affinity measure

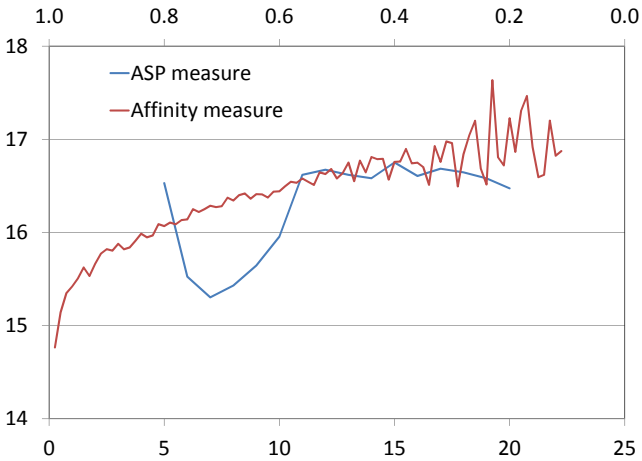

Figure 8: Average age difference between pairs of users as a function of ASP or affinity measure

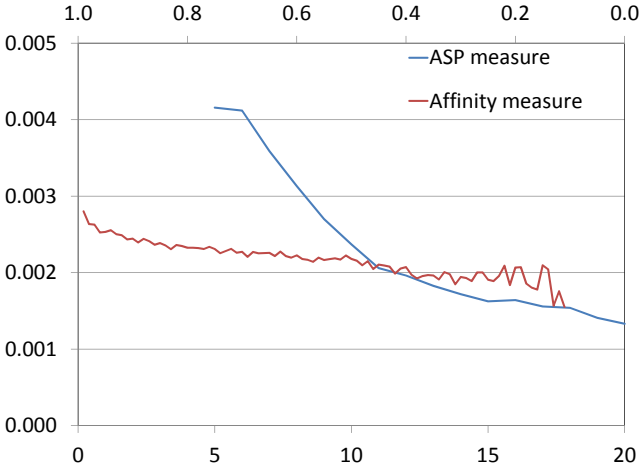

Figure 9: Average cosine similarity of queries of users as a function of ASP or affinity measure

	negative ASP distance	A^θ unweighted	A unweighted	A^θ weighted	A weighted	A^θ log-weighted	A log-weighted
Country similarity	0.0828	0.0020	0.0020	0.0255	0.0258	0.0179	0.0181
Language similarity	0.1794	0.0324	0.0329	0.0430	0.0439	0.0424	0.0432
Age similarity	-0.0290	0.0241	0.0242	-0.0110	-0.0110	-0.0015	-0.0011
Query similarity	0.0616	0.0200	0.0205	0.0155	0.0161	0.0186	0.0191

Table 1: Correlation coefficient between similarity of pairs of users and their proximity in the email exchange graph.

Figure 4 shows the distribution of ASP distances. The mode of the distribution is at distance 10, and the bulk is between 5 and 15. It should be pointed out that in earlier work we found ASP to overestimate distance by a constant additive amount (about 2 to 3 in the case of undirected web graphs), so we assume the distribution of true distances to be shifted accordingly. Figure 5 shows the distribution of A^θ affinity values (for $\theta = 0.5$) of the sketches for the graph with unweighted edges. The mode of the distribution is at 0.82. Given that the sampling probability vector $p_1, \cdots, p_r$ was calibrated to produce a uniform distribution of affinity values across the entire set of 312 million users in the email exchange graph, the non-uniformity of the distribution in Figure 5 indicates that our set of candidate users is biased towards users with affinity towards each other.

Figure 6 shows the fraction of pairs of users (drawn from the candidate set) residing in the same country, as a function of the distance and affinity between the pair of users. The vertical axis shows the fraction of pairs coming from the same country; the bottom scale of the horizontal axis shows ASP distance values (in increasing order) whose distribution is plotted by the blue curve; the top scale shows the estimated A^θ affinity values (in decreasing order, for $\theta = 0.5$ and based on the sketches formed using unweighted edges) whose distribution is plotted by the red curve. Each (x, y) point of a curve shows the fraction y of all pairs at distance/affinity x who reside in the same country. In order to reduce noise, we restricted the curves to only those distance and affinity values with at least 1000 (out of 5 million) pairs of users. Both distance and affinity measures are discrete, but since the set of possible distance values is smaller than the set of possible affinity values, the affinity curve appears smoother. The ASP curve is far more slanted than the affinity curve, suggesting that the correlation between ASP and country of residence is far stronger than between affinity and country. Indeed, this is borne out by the correlation coefficients shown in Table 1.

Figure 7 shows the fraction of pairs of users speaking the same language, using the same encoding as the previous figure. Again, we restricted the curve to points with at least 1000 pairs of users. As in the previous figure, the ASP curve is far more slanted than the affinity curve, suggesting stronger correlation between distance and language than between affinity and language. This is also borne out by the correlation coefficients in the second row of Table 1.

Figure 8 shows the age difference between pairs of users as a function of the distance and affinity between the pair of users. The vertical axis plots the age difference, the horizontal axis plots distance and affinity as it did in Figures 6 and 7. Each (x, y) point of a curve shows the average age difference y of all pairs at distance/affinity x. Again, the curve only shows points with at least 1000 pairs of users so as to reduce noise. The graph has several features worth pointing out: First, note that the vertical axis does not start at 0, but rather at 14 years – the average age difference between pairs of users is ranges between about 15 and 18 years regardless of distance or affinity. Putting it differently, email is a social medium that connects generations. Second, age difference and affinity are negatively correlated – users with high affinity are on average closer in age. Third, the distance curve is non-monotonic – there is no clear relationship between the ASP distance of two users in the email graph and their age difference. Referring to the third row of Table 1, the correlation coefficients of ASP distance and age vs. affinity and age are quite low and fairly close to each other, with ASP having a slight edge.

The previous three figures examined the connection between two users' demographic background (namely country, language and age) and their distance and affinity in the email exchange graph. Next, we examine whether the distance or affinity between two users is predictive of shared *interests*. To that end, we aggregate all the queries issued by a given user into a bag of terms, and we weigh each term in the bag by its tf-idf value [21]. We compute the cosine similarity between the tf-idf weighted queries for each of the 5 million pairs of users, and we treat this similarity as a proxy of the similarity of their interests. Figure 9 depicts the results of this experiment. The vertical axis plots the similarity between a pair's queries, the horizontal axis plots distance and affinity as it did in the previous three figures. Each (x, y) point of a curve shows the average similarity y of all pairs at distance/affinity x. As in Figures 6 and 7, the ASP curve is more slanted than the affinity curve, suggesting stronger correlation between distance and query similarity than between affinity and query similarity. This is also borne out by the correlation coefficients in the fourth row of Table 1.

Table 1 shows the correlation coefficient between user demographic and interest profiles and the different types of social proximity measures we are considering in this study. In addition to the approximate shortest path distance, we are considering six different variants of the affinity measure: A^θ (for $\theta = 0.5$) and A, computed on the email exchange graph with unweighted edges, edges weighted (and sampled) by the number of reciprocal email exchanges, and edges weighted by the logarithm of that number. Consistent with Figures 6–9, we find that the correlation coefficients of the ASP measure are higher than those of A^θ and A for all user characteristics. The weighted versions of affinity also have a higher correlation than the unweighted ones for all characteristics except age. This might indicate that age changes differently over the network than language or country. Language and country probably change over long distances, whereas age per-

	negative ASP distance	A^θ unweighted	A unweighted	A^θ weighted	A weighted	A^θ log-weighted	A log-weighted
Country similarity	75.27	6.67	10.00	15.00	15.18	16.27	15.08
Language similarity	81.55	13.50	13.71	12.29	12.19	11.46	11.68
Age similarity	13.81	7.09	7.33	7.86	8.46	15.00	11.00
Query similarity	23.69	6.90	7.32	10.33	9.47	7.15	7.64

Table 2: Statistical significance of the correlations: Ratio of correlation for the same pair to correlation for random pairs.

haps changes more quickly; that is, the ball around a point with similar age is probably smaller than the ball around a point with the same language or country. This could be an indication that the unweighted variants of affinity are better measures to capture "small" social distances whereas ASP is better for "large" social distances. This is consistent with our earlier observation that ASP has a higher percentage error on shorter distances [11]. It is an open question of how well *precise* shortest-path distance would perform; unfortunately we do not know how to perform five million precise shortest-path computations within a reasonable time frame.

While it may seem that the correlation coefficients in Table 1 are small, we show that they are much higher than the values we would expect if the properties were completely uncorrelated. In Table 2, we compare the correlation coefficients computed in Table 1 to the correlation coefficients between the same properties for two independently chosen sequences of 5 million pairs. For example, to compute the correlation coefficient between affinity and cosine similarity ($C_{A,cs}$), we computed it between the vector of cosine similarities and affinities on the same set of 5 million pairs. Now we compute it between the vector of cosine similarities for one random set of pairs and vector of affinities for another random set of pairs ($C^r_{A,cs}$). We compute the ratio $|C_{A,cs}|/|C^r_{A,cs}|$. If this ratio is high it means that the values of affinity and cosine similarities for the same pair are much more correlated than for two random pairs. Table 2 shows this ratio for the different pairs of properties. As can be seen, the ratio is about 7 to 16 for the affinity variants and about 13 to 80 for the ASP measure, which means that the statistical significance of the correlations is not negligible.

3.2 The Twitter experiment

Our second experiment is based on one month worth of Twitter postings (*tweets*), 1,475,522,405 in total. Twitter is a popular micro-blogging service that allows users to post messages up to 140 characters in length. The Twitter community follows a number of stylistic conventions when composing tweets, one of them the convention of mentioning other Twitter users by prefixing their Twitter user IDs with the @ character. A single *mention* in itself does not indicate a social connection between the mentioner and the mentioned, but we assert that if two users mutually mention each other, a social bond does indeed exist.

We first ran an initial data extraction process on our collection of tweets that identified all pairs of (distinct) users mutually mentioning each other at least once during August 2011, and extracted all tweets (whether containing mentions or not) authored by these users. This extracted 1,265,660,845 tweets authored by 10,410,144 users.

Next, we constructed the mention-graph using the following process: First, we constructed an intermediate weighted directed graph containing a vertex for each user whose tweets we extracted during the data extraction phase, and an edge (u, v) with weight w if user u mentioning user v a total of w times. Next, we constructed the actual mention graph, an undirected weighted graph, by taking the vertex set of the directed graph, and introducing an edge between u and v with weight $mean(w_1, w_2)$ if the directed graph contained an edge (u, v) with positive weight w_1 and an edge (v, u) with positive weight w_2. We constructed one variant of the mention graph using the arithmetic mean of weights, and another variant using the geometric mean. The mention-graph contained 10,410,144 vertices and 176,551,621 edges.

As in the previous experiment, we drew a random sample of 5 million pairs of users, and computed affinity values and ASP distances between each pair of users in the mention graph. We also computed the cosine similarity between the tf-idf vectors of the collected tweets of each pair of users, and we computed correlation coefficients between the graph-based measures and the text-based ground truth. 880,564 of the 5,000,000 pairs (or 17.6% of all sampled pairs) were in separate connected components of the mention-graph, with no path between them. These pairs have an affinity value of 0 and an ASP distance of ∞.

In computing affinity values and ASP distances, we experimented with different transform functions for edge weights. In the case of affinity, where heavier edges are more likely to be included in G_p and thus denote closeness, we tried the constant transform $T(w) = 1$ (effectively ignoring weights), the identity transform $T(w) = w$ (directly using the mean of the mention frequency), and the logarithmic transform $T(w) = \log w$ (discounting excessive "mutual admiration"). In the case of ASP distance, where edge weight denote social distance, we tried the constant transform $T(w) = 1$ (effectively ignoring weights), the reciprocal transform $T(w) = \frac{1}{w}$, and the reciprocal logarithmic transform $T(w) = \frac{1}{\log w}$.

Figure 10 shows the distribution of ASP distances computed using the constant edge weight transform (i.e. an unweighted graph); Figure 11 shows the distribution of A^θ affinity values (for $\theta = 0.5$) of the sketches for the graph with unweighted edges. In both figures, we do not show the 880,564 disconnected pairs of users in the sample. The mode of the ASP distance distribution is at distance 9, and the bulk is between 6 and 12.

Figure 12 is the analogous version of Figure 9 for the Twitter mention graph. We compute the cosine similarity between the tf-idf weighted tweets of each of the 5 million pairs of users, and we treat this similarity as a proxy of the similarity of their interests. The vertical axis plots the similarity between a pair's queries, the horizontal axis plots distance

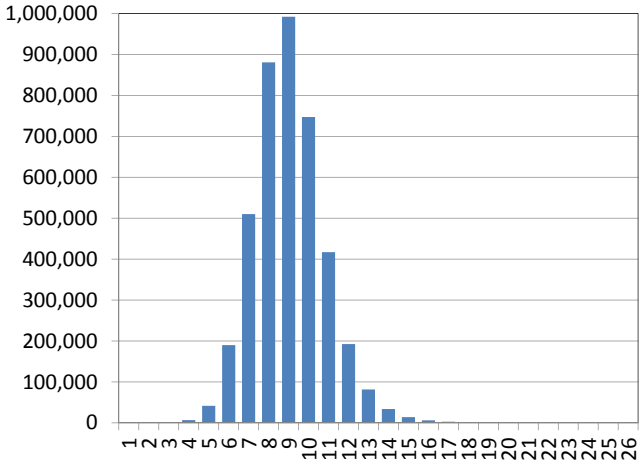

Figure 10: Distribution of ASP distances for the sampled 5 million pairs of users in the Twitter graph

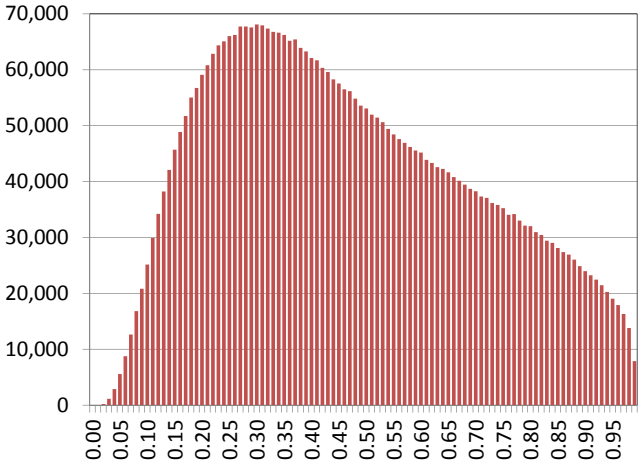

Figure 11: Distribution of affinities for the sampled 5 million pairs of users in the Twitter graph

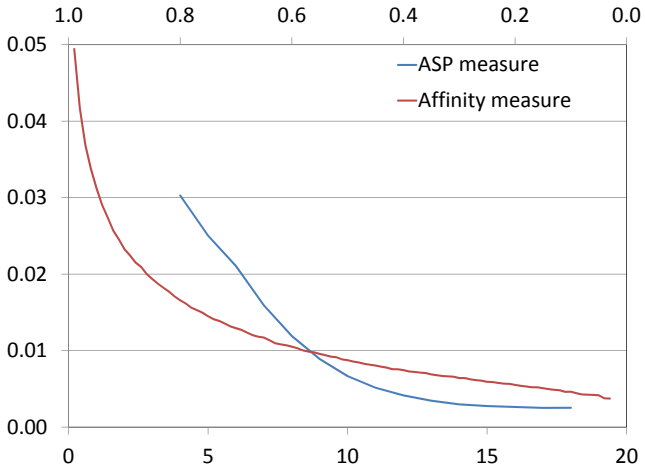

Figure 12: Average cosine similarity of queries of users as a function of ASP or affinity measure

and affinity. Each (x, y) point of a curve shows the average similarity y of all pairs at distance/affinity x. Unlike what we observed for the Hotmail graph, the affinity curve is more slanted than the ASP curve, suggesting stronger correlation between affinity and tweet similarity than between distance and tweet similarity. This is also borne out by the correlation coefficients in Table 3. The table shows correlation coefficients between textual cosine similarity (our ground truth) and different variants of our graph-distance measures. Specifically, it explores the impact of using an arithmetic vs. a geometric mean when combining edge weights during the mapping from a directed to an undirected graph; the impact of multiple edge weight transforms; and of course the distance measure itself – negated ASP distance, A^θ affinity, and A affinity. We can draw several conclusions from the data presented in this table:

- Similarity is much more highly correlated to affinity than to ASP distance.
- For affinity measures, the identity transform performs slightly better than the other transforms.

	arithmetic mean	geometric mean
ASP, $T(w) = 1$	0.216	0.216
ASP, $T(w) = \frac{1}{w}$	0.206	0.207
ASP, $T(w) = \frac{1}{\log w}$	0.209	0.210
A^θ, $T(w) = 1$	0.472	0.468
A^θ, $T(w) = w$	0.489	0.487
A^θ, $T(w) = \log w$	0.488	0.484
A, $T(w) = 1$	0.472	0.469
A, $T(w) = w$	0.492	0.490
A, $T(w) = \log w$	0.489	0.486

Table 3: Correlation coefficient between similarity of pairs of users and their proximity in the Twitter mention graph.

- In the context of the affinity measures, combining edge weights using their arithmetic mean is slightly better than using their geometric mean.
- The A affinity measure performs slightly better than the A^θ affinity measure.

4. CONCLUSION

In this paper, we introduced "social affinity" as a new measure of how robustly two users are connected within a social network. Intuitively, affinity captures how robust the connection between a pair of users is to random deletion of edges in the graph. The affinity value can be efficiently estimated through a sketch-based algorithm, which requires only two table lookups to retrieve the sketch for each user, plus a simple computation on the pair of sketches. We compare this new affinity measure against "approximate shortest path" distance, another sketch-based measure estimating the shortest-path distance between two nodes in a graph. We evaluated our measure on two social graphs. First, we used a sampling of the Hotmail email exchange graph as our social network, and user profiles and query history as independent measures of user similarity. We found that both ASP distance and social affinity between two users are correlated to their profile and query similarity. While ASP had a higher correlation for most user properties, affinity was bet-

ter in one that changes over short distances. This suggests that the affinity measure may be better-suited for capturing short-distance effects in a social graph, whereas ASP may be preferable for capturing long-distance effects. Second, we constructed a Twitter mention graph from one month worth of tweets, and the collection of tweets written by a given user as the ground truth. We found that the textual similarity between two users' tweets was strongly correlated to their affinity, and somewhat more weakly correlated to their approximate distance in the mention-graph.

We hypothesize that some of the weaknesses we observed for ASP are due to the fact that ASP is an approximation of the true distance. We are still investigating whether it is reasonably feasible to compute *precise* shortest path distances for the five million pairs we used to compare our measures.

5. REFERENCES

[1] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. In *41st Annual ACM Symposium on Theory of Computing*, 2009.

[2] András A. Benczúr and David R. Karger Approximating $s-t$ minimum cuts in $\tilde{O}(n^2)$ time. In *28th Annual ACM Symposium on the Theory of Computing*, 1996.

[3] Jean Bourgain. On Lipschitz embeddings of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52(1-2):46–52 (1985).

[4] Andrei Z. Broder. Identifying and filtering near-duplicate documents. In *11th Annual Symposium on Combinatorial Pattern Matching*, 2000.

[5] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the web. *Computer Networks*, 29(8):1157–1166 (1997).

[6] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* 357(4): 370-379 (July 2007).

[7] Nicholas A. Christakis and James H. Fowler. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 358(21): 2249-2258 (May 2008).

[8] David J. Crandall, Dan Cosley, Daniel P. Huttenlocher, Jon M. Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[9] Dan Cosley, Daniel P. Huttenlocher, Jon M. Kleinberg, Xiangyang Lan, and Siddharth Suri. Sequential influence models in social networks. In *4th International Conference on Weblogs and Social Media*, 2010.

[10] Pedro Domingos and Matthew Richardson. Mining the network value of customers. In *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[11] Atish Das Sarma, Sreenivas Gollapudi, Marc Najork, and Rina Panigrahy. A sketch-based distance oracle for Web-scale graphs. In *3rd International Conference on Web Search and Web Data Mining*, 2010.

[12] James H. Fowler and Nicholas A. Christakis. The dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal* 2008; 337: a2338.

[13] Wai Shing Fung, Ramesh Hariharan, Nicholas J. A. Harvey, and Debmalya Panigrahi. A general framework for graph sparsification. In *43rd ACM Symposium on Theory of Computing*, 2011.

[14] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[15] Gueorgi Kossinets, Jon M. Kleinberg, and Duncan J. Watts. The structure of information pathways in a social communication network. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[16] Ioannis Koutis, Gary L. Miller, and Richard Peng. Approaching optimality for solving SDD linear systems. In *51st Annual IEEE Symposium on Foundations of Computer Science*, 2010.

[17] Jure Leskovec, Ajit Singh, and Jon M. Kleinberg. Patterns of influence in a recommendation network. In *10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2006.

[18] Ulrike von Luxburg, Agnes Radl, Matthias Hein. Getting lost in space: Large sample analysis of the commute distance. In *24th Annual Conference on Neural Information Processing Systems*, 2010.

[19] Jiri Matousek. On the distortion required for embedding finite metric spaces into normed spaces. *Israel Journal of Mathematics*, 93(1):333–344 (1996).

[20] J. Niels Rosenquist, Joanne Murabito, James H. Fowler, and Nicholas A. Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine* 152(7): 426-433 (April 2010).

[21] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24 (5): 513–523 (1988).

[22] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *40th Annual ACM Symposium on Theory of Computing*, 2008.

[23] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *36th Annual ACM Symposium on Theory of Computing*, 2004.

[24] Mikkel Thorup and Uri Zwick. Approximate distance oracles. *Journal of the ACM*, 52(1):1–24 (January 2005).