Computer Communications 73 (2016) 188–199

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

The power of indirect ties

Xiang Zuo[a,*], Jeremy Blackburn[b], Nicolas Kourtellis[b], John Skvoretz[c], Adriana Iamnitchi[a]

[a] Computer Science and Engineering, University of South Florida, FL, 33620, USA
[b] Telefonica Research, Barcelona, Spain
[c] Department of Sociology, University of South Florida, FL 33620, USA

ARTICLE INFO

Article history:
Available online 16 July 2015

Keywords:
Social strength
Indirect tie
Network dynamic
Information diffusion path
Friend-to-friend storage systems

ABSTRACT

While direct social ties have been intensely studied in the context of computer-mediated social networks, indirect ties (e.g., friends of friends) have seen little attention. Yet in real life, we often rely on friends of our friends for recommendations (of good doctors, good schools, or good babysitters), for introduction to a new job opportunity, and for many other occasional needs. In this work we attempt to 1) quantify the strength of indirect social ties, 2) validate the quantification, and 3) empirically demonstrate its usefulness for applications on two examples. We quantify social strength of indirect ties using a measure of the strength of the direct ties that connect two people and the intuition provided by the sociology literature. We evaluate the proposed metric by framing it as a link prediction problem and experimentally demonstrate that our metric accurately (up to 87.2%) predicts link's formation. We show via data-driven experiments that the proposed metric for social strength can be used successfully for social applications. Specifically, we show that it can be used for predicting the effects of information diffusion with an accuracy of up to 0.753. We also show that it alleviates known problems in friend-to-friend storage systems by addressing two previously documented shortcomings: reduced set of storage candidates and data availability correlations.

1. Introduction

Mining the huge corpus of social data now available in digital format has led to significant advances of our understanding of social relationships and behavior [1] and confirmed long standing results from sociology on larger datasets. In addition, social information (mainly relating people via declared relationships on online social networks or via computer-mediated interactions) has been successfully used for a variety of applications, from spam filtering [2] to recommendations [3] and peer-to-peer backup systems [4].

All these efforts, however, focused mainly on direct ties. Direct social ties (that is, who is directly connected to whom in the social graph) are natural to observe and reasonably easy to classify as strong or weak [5,6]. Indirect social ties, though, defined as a relationship between two individuals who have no direct relation but are connected through a third person in their social network [7], carry a significantly larger potential as they facilitate better information dissemination then direct ties [8] and enable significantly better opportunities [9]. Computer-mediated applications, we conjecture, have a significantly higher potential in mining and exploiting indirect ties, as the direct ties are likely to be used via the traditional channels through which were used for thousands of years.

However, not all indirect ties are valuable or useful, even at short distances (i.e., 2 hops). For example, a distant acquaintance of a mere acquaintance is unlikely to have a social incentive for performing a personal favor, such as sharing available storage on his personal computer. Moreover, trust is likely diluted under such conditions. Why would a weak distant social contact trust that the data he is asked to store is not illegal or malicious? In addition, what works for a user or an application might not work for another user or another application: the indirect tie $A - X - B$ may be strong enough for A to use, but not enough for B to use; or it may be strong enough to use for a backup application, but not for a social contagion. Therefore, quantifying the strength of an indirect tie is both necessary and non-trivial.

In this paper, we build upon and further adapt a metric called *social strength*, which we introduced in [10,11], that quantitatively estimates the strength of an *indirect* tie. Our metric uses various observations from sociology and builds on the current opportunities of quantifying the strength of direct ties from computer or phone-recorded interactions. We rely on the sociology literature to define the requirements of such a metric (Section 2). First, since social relationships are asymmetrically reciprocal [12], the social strength of an indirect tie consequently needs to be asymmetrical as well. Second, a friend of many of one's friends — thus connected via multiple 2-hop paths — can potentially be more socially "close" than the friend

[*] Corresponding author. Tel.: +18137487106.
E-mail address: xiangzuo@mail.usf.edu, xiangzuo2012@gmail.com (X. Zuo).

http://dx.doi.org/10.1016/j.comcom.2015.07.013
0140-3664/© 2015 Elsevier B.V. All rights reserved.

of a friend, connected via only one 2-hop path. Third, the strength of an indirect tie decreases with the length of the shortest path [13]. In Section 4 we validate the social strength metric using real datasets.

We demonstrate the usefulness of our metric on two proof-of-concept applications. First, in Section 5, we show that the social strength metric can be used for inferring, and in effect, predicting information diffusion paths, which further implies the influence of indirect ties on information flow in network dynamics. Second, in Section 6, we experimentally show that two main issues identified in friend-to-friend storage systems, namely reduced candidate sets [4] and low availability due to time synchronization among friends [14], are significantly alleviated by employing our social strength metric for the recruitment of socially close indirect contacts as storage candidates. We discuss our findings and conclude in Section 8.

2. Social strength definition

We want to define a metric that quantifies the strength of a social connection between indirectly connected nodes in a social network. The need for such a metric is supported by many sociological studies and is also intuitively understood from daily life: friends of friends are an important resource for information and useful social contacts.

In our attempt to quantify an indirect social tie, we use the following observations from sociology and from recent data-driven studies on computer-mediated social relationships:

O1: The strength of a direct social relationship is related to the amount of interactions, as shown in [8,15]: the more frequently persons interact with one another, the more likely they will form strong relationships. Moreover, interactions among OSN users were shown to represent more meaningful relations than just declared relationships [16]. Consequently, in the quantification of an *indirect* social tie, we rely on a numerical representation of the strength of a *direct* social tie that can be expressed as number of interactions, number of shared interests, or other recordable outcomes, depending on the semantics of the relationship.

O2: The strength of an indirect tie decreases with the length of shortest path between the two individuals. This has been quantitatively observed by Friedkin [13], who concluded that people's awareness of others' performance decreases beyond 2 hops. Three degrees of influence theory, proposed by Christakis et al. [17] states that social influence does not end with people who are directly connected but also continues to 2- and 3-hop relationships, albeit with diminishing returns. This theory has held true in various social networks examined [18,19]. In accordance with these observations, the social strength metric we propose focuses on 2- and 3-hop relationships with a decreasing value as a function of distance.

O3: Multiple types of social interactions (for example, both professional collaboration and playing tennis after work) result into a stronger (direct) relationship than only one type of interaction [20]. Furthermore, sociology studies [13] observed that the relationship strength of indirectly connected individuals greatly depends on the number of different direct or indirect paths connecting them. Therefore, we consider the strength of multiple shortest paths in our definition of the strength of an indirect social tie.

O4: Typically, social ties between individuals are asymmetrically reciprocal [21]. Thus, for the directly connected users Alice and Bob, the importance of their mutual relationship may be dramatically different. We want to preserve this asymmetry in quantifying indirect ties, such that Alice and Charlie, indirectly connected via Bob, are entitled to have different views about their indirect tie.

Therefore, to quantify the social strength of an indirect social tie between users i and m, we consider relationships at n social hops ($n = 2$ or 3), where n is the shortest path between i and m. We assume a weighted interaction graph model that connects users with edges weighted based on any type of signal (information) that can represent tie strength of their direct relationships. Assuming that $\mathcal{P}^n_{i,m}$ is the set of different shortest paths of length n joining two indirectly connected users i and m and $\mathcal{N}(p)$ is the set of nodes on the shortest path p, $p \in \mathcal{P}^n_{i,m}$, we define the social strength between i and m from i's perspective over an n-hop shortest path as:

$$SS_n(i,m) = 1 - \prod_{p \in \mathcal{P}^n_{i,m}} \left(1 - \frac{\min\limits_{j,\ldots,k \in \mathcal{N}(p)} [NW(i,j),\ldots,NW(k,m)]}{n}\right) \tag{1}$$

This definition uses the normalized direct social weight $NW(i, j)$ between two directly connected users i and j, defined as follows:

$$NW(i,j) = \frac{\sum_{\forall \lambda \in \Lambda_{i,j}} \omega(i,j,\lambda)}{\sum_{\forall k \in N_i} \sum_{\forall \lambda \in \Lambda_{i,k}} \omega(i,k,\lambda)} \tag{2}$$

$NW(i, j)$ calculates the strength of a direct relationship by considering all types of interactions $\lambda \in \Lambda$ between the users i and j such as phone calls, interactions in online games or number of co-authored papers (observation O3). These interactions are normalized to the total amount of interactions of type λ that i has with other individuals. This approach ensures the asymmetry of social weight (observation O4) in two ways: first, it captures the cases where $\omega(i, j, \lambda) \neq \omega(j, i, \lambda)$ (such as in a phone call graph). Second, by normalizing to the number of interactions within one's own social circle, even in undirected social graphs, the relative weight of the mutual tie will be different from the perspective of each user.

The observations O1, O3 and O4 were incorporated in the definition of the NW function and naturally carry over in the definition of social strength from $SS_n(i, m)$. Moreover, O3 is additionally implemented by considering the product over all shortest paths p that connect two users. O2 is implemented by considering the weakest link (minimum normalized weight of all direct ties on each path) and by dividing it with the distance n between the users. The proposed social strength measure can:

- Quantify the indirect tie strength for nodes indirectly connected at any social distance.
- Treat indirect ties between two nodes as possibly asymmetric in strength rather than constraining the values to be equal.
- Be more sensitive to strength differences because it uses both edge weights and number of paths to calculate a value.
- Be calculated without graph's global information.

3. Datasets

In this paper we use several datasets from different domains. Our datasets vary from fast, non-profound dynamics to slow professional networks and more traditional social networks augmented with heavy interactions.

Team fortress 2 (TF2) is an objective-oriented first person shooter game released in 2007. We collected more than 10 months of gameplay interactions (from April 1, 2011 to February 3, 2012) on a TF2 server [22]. The dataset includes game-based interactions among players, timestamp information of each interaction, declared relationship in the associated gaming OSN, Steam Community [23], and the time when the declared friendship was recorded. The resulting TF2 network is thus composed of edges between players who had at least one in-game interaction while playing together on this particular server, and also have a declared friendship in Steam Community. This dataset has several advantages over the Steam declared OSN:

Table 1
Characteristics of the social networks used in our experiments. APL: average path length, CC: clustering coefficient, EW: range of edge weights, OT: observation time.

Networks	Nodes	Edges	APL	Density	CC	Assortativity	Diameter	EW	OT
TF2	2,406	9,720	4.2	0.0034	0.21	0.028	12	[1–21,767]	300 days
IE	410	2,765	3.6	0.0330	0.45	0.225	9	[1–191]	90 days
CA-I	348	595	6.1	0.0098	0.28	0.173	14	[1–52]	N/A
CA-II	1,127	6,690	3.4	0.0100	0.33	0.211	11	[1–127]	N/A

First, it provides the number of in-game interactions that can be used to quantify the strength of a social tie. Second, it provides players' online/offline status that we use later in the experiments in Section 6. Third, each interaction and friendship formation is annotated with a timestamp, which is helpful for examining the dynamics of links under formation. Fourth, over a pure in-game interaction network, it has the advantage of selecting the most representative social ties, as proven in [22]. In this network of 2.4k nodes and 9.7k edges, edge weights represent the number of in-game interactions.

Infectious exhibition (IE) held at the Science Gallery in Dublin, Ireland, from April 17th to July 17th in 2009 was an event where participants explored the mechanisms behind contagion and its containment. Data were collected via radio-frequency identification (RFID) devices that recorded face-to-face proximity relations of individuals wearing badges [24]. Each interaction was annotated with a timestamp. We translated the number of interactions into edge weights.

Co-authorship networks (CA-I and CA-II) are extracted from ArnetMiner[1] and are based on papers co-authored by Computer Science researchers [25]. Nodes in these graphs represent authors and edges between two nodes are weighted with the number of papers co-authored by the two authors. From this dataset we extracted the two largest connected components (see Table 1 for details). *Co-authorship I (CA-I)* is a small connected component and a relatively low density. *Co-authorship II (CA-II)* is the largest connected component of the ArnetMiner co-authorship network, having a density one order of magnitude higher than CA-I. Because the dataset does not include time publication information, the observation window is unspecified in Table 1.

A brief characterization of the networks appears in Table 1. Fig. 1 plots the degree, edge weight, and clustering coefficient distributions for each network. We note that IE is a smaller but much denser network, while TF2's interactions frequency is much higher than the other networks', as shown by the range of edge weights. Even though CA-I and CA-II are extracted from the same OSNs, they have different degree and clustering coefficient distributions. Since they contain timestamps of the links formed and interactions between users, we use TF2 and IE networks to validate our proposed social strength metric by studying link formation in Section 4. We use the TF2 and CA networks to study diffusion and peer expansion in Sections 5 and 6, respectively, as they are larger, sparser and based on longer lasting relationships compared to IE's ad-hoc interactions.

4. Social strength evaluation

In sociology, the theory of homophily [26] postulates that people tend to form ties with others who have similar characteristics. Moreover, a stronger relationship implies greater similarity [8]. Therefore, a number of link prediction models that estimate tie strength from graph structure [27] or interaction frequency and users' declared profiles similarities [28] have been proposed.

To verify that social strength is in fact quantifying the strength of social ties, we frame it as a link prediction problem. Simply put, given a pair of users, the link prediction problem asks whether the strength of the tie is strong enough to form a social relationship between them. Specifically, given a snapshot of a social network, we use social strength values to infer which new relationships or interactions among users are likely to occur in the near future. Granovetter put forth the idea of the "forbidden triad" [8], i.e., a triad where there is a strong tie between say u and v as well as a strong tie between v and w, but no tie between u and w is extremely unlikely to exist. Under the theory of triadic closure, forbidden triads will quickly close because a relationship will form between u and w. Thus, if we can effectively predict edge formations based on the value of social strength, the implication is that social strength is capturing the strength of ties between distant nodes. We compare our results with three other metrics used for link prediction.

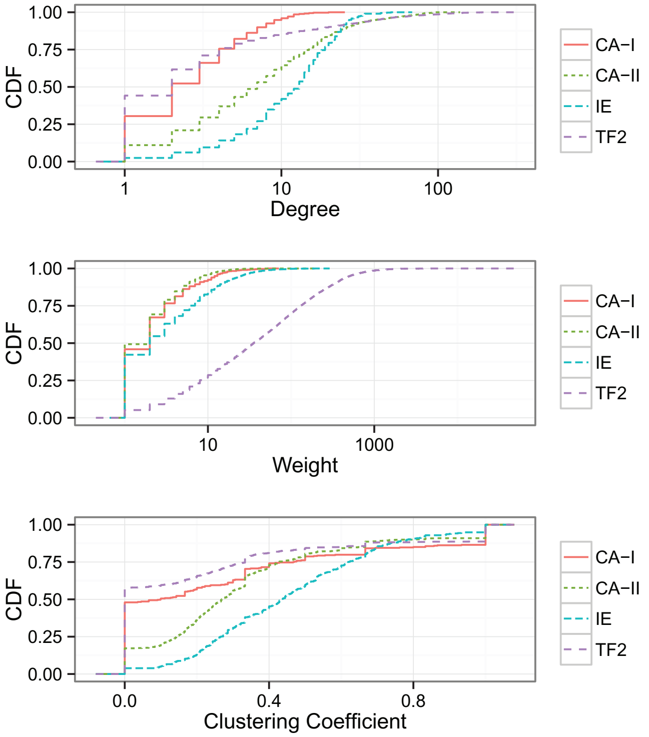

Fig. 1. CDF of the degree, weight and clustering coefficient for the four networks used.

4.1. Compared metrics

We compare three well-established link prediction metrics with the social strength metric to demonstrate how effective is in link prediction. Many approaches are based on the idea that if two nodes i and j have large overlap in their neighbors, they have higher likelihood to form a link in the future. In the following definition, let $\Gamma(i)$ denote the set of node i's neighbors.

Jaccard coefficient (J) is a commonly used similarity metric that was proposed by Salton and McGill [29]:

$$J(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}$$

[1] http://arnetminer.org/

Adamic-Adar (AA) is a metric that only counts common features by inverting log frequency of their occurrence [30]:

$$AA(i, j) = \sum_{z\in\Gamma(i)\cap\Gamma(j)} \frac{1}{log|\Gamma(z)|}$$

Katz defined a metric that sums all possible paths between two nodes [31]:

$$Katz(i, j) = \sum_{l=1}^{\infty} \beta^{l} \cdot |P^{l}_{i,j}|$$

$P^{l}_{i,j}$ is the set of all $length - l$ paths between i and j. Paths are exponentially damped by length, so that shorter paths count more heavily. β ($\beta > 0$) is a parameter that if set at a very small number, the measurement is similar to the *common neighbors* metric that directly counts the common friends between two nodes, since more than 2-hop path lengths contribute little to the summation. According to the experimental results in [32] (Fig. 3), $\beta = 0.0005$ gives better prediction performance than other values. Thus, in the following experiments we set the parameter β to 0.0005. Note that even if a weighted Katz metric is discussed in [32], it is only applicable to 1-hop social distance that is not suitable to our problem (2- and 3-hop distance), thus, in this paper, we only conduct comparisons with unweighted Katz metric as described below.

All these metrics give a score that quantifies the strength of the social tie between two nodes. Jaccard and Adamic-Adar are based on node neighborhoods while Katz uses the ensemble of all paths between two nodes. Therefore, Jaccard and Adamic-Adar restrict their measurements to nodes that are 2-hops away while Katz can be applied to n-hop ($n \geq 1$) social distance, which is comparable to our social strength metric on longer path lengths.

4.2. Experiments

As we explained in Section 2, people can be aware of others' behaviors within 2 hops and be influenced by indirect ties up to 3 hops. Thus, we focus our experiments on 2- and 3-hop social distance.

4.2.1. Experimental setup

We formalize the link prediction task as a binary classification problem that predicts whether an edge exists in the graph or not, when two users are 2- or 3-hops away from each other. First, for all pairs of users u and v, we label them with the timestamp of their relationship (t_{uv}), or $\emptyset$ in the case that there was no relationship by the end of our observation period. Next, we compute the 2- and 3-hop social strength between each u, v pair based on the state of the graph at time $t_{uv} - 1$, or in the case of $\emptyset$, the final state of the graph.

For example, Fig. 2a indicates the state of the subgraph of nodes A, B, and C at the end of the observation period. When given the pair B, C, whose relationship formed at t_3, the calculation of their social strength includes the edges AB and AC since they both formed prior to t_3. Conversely, when given the pair A, B, none of the edges are included in the social strength calculation. A similar scenario of a 3-hop prediction is depicted in Fig. 2c.

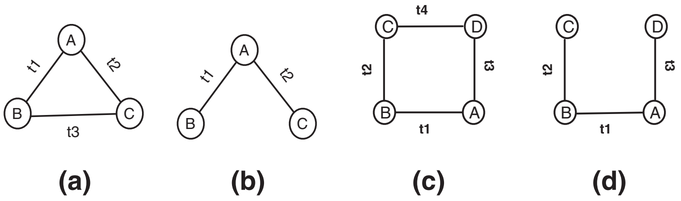

Fig. 2. Demonstration of a pair of nodes' relationship status: (a) a pair of nodes, B and C, have a relationship in 2-hop social distance before t_3, where $t_1 < t_3$ and $t_2 < t_3$, and they formed a direct relation at t_3. (b) a pair of nodes, B and C, have a relationship in 2-hop social distance and no direct relation formed ($\emptyset$) by the end of our observation period. (c) a pair of nodes, C and D, have a relationship in 3-hop social distance before t_4, where $t_1 < t_4$, $t_2 < t_4$, and $t_3 < t_4$, and they formed a direct relation at t_4. (d) a pair of nodes, C and D, have a relationship in 3-hop social distance and no direct relation formed ($\emptyset$) by the end of our observation period.

The TF2 network has a timestamp of when a declared relationship was created, but the IE network only has the timestamp of the first recorded face-to-face interaction between two individuals. Thus, for IE, we use this timestamp as a proxy for the creation of a relationship.

For 2-hop social distance, there are 5, 984 pairs in TF2 with $t_{uv} \neq \emptyset$, i.e., that had a relationship form prior to the end of the observation period, and 161,561 pairs with $t_{uv} = \emptyset$; 2,475 with $t_{uv} \neq \emptyset$ and 676, 863 with $t_{uv} = \emptyset$ for 3-hop distant users. IE has 1,886 formed relations and 4,111 unformed for 2 hops, and 484 formed relations and 24, 631 unformed for 3 hops. In other words, our datasets are imbalanced with respect to formed relationships and no relationships ($\emptyset$). There are two common approaches for dealing with imbalanced data classifications: under-sampling [33] and over-sampling [34]. We chose to under-sample users with no relationships, thus in our experiment they appear at the same empirical frequency as the formed relationships.

4.2.2. Results

In our prediction tasks, we use a classic tree-structured machine learning classifier, *Decision Tree* (J48), and the scores calculated from evaluation metrics as features. Note that because we treat social relationships asymmetrically, social strength outputs two different scores, each coming from the node's own perspective. We compare the performance of our social strength to the three tie strength metrics introduced in Section 4.1. Four evaluators — Precision, Recall, F-Measure and Area Under Curve (AUC) — are used to evaluate the prediction performance. Table 2 shows the link prediction results of nodes 2 hops away. We run 10-fold cross validation [35]. Among all evaluations and classifiers, social strength outperforms other metrics in predicting link existence between pairs of users. We note that the AUC arrives up to 0.765 for the TF2 network and reaches 0.872 for the IE network when using social strength, greatly outperforming the other predictor metrics.

Next, we test the social strength viability at predicting link existence from 3-hops away as people are influenced by others within 3-hop social distance [17]. However, both Jaccard and Adamic-Adar metrics are restricted to predictions within 2 hops, thus, only social strength and Katz results appear in Table 2. Social strength outperforms Katz might relate to two factors in theory. First, social strength considers edge weights and values ties asymmetrically. Second, in Katz, the same parameter is set for every pair of nodes, which is not accurate to capture the difference of tie strength among different pairwise nodes. We note that while social strength's effectiveness is reduced, it still manages to properly discriminate between existing and non-existing links up to ~64.5% of the time in TF2 and 67.1% of the time in IE. While it is expected to see a decrease in performance when we cross the horizon of observability of 2 hops [13], our results show that social strength preserves a quantification of the strength of *indirect* social ties.

4.3. Relevant information for measuring indirect ties

As Table 2 shows, the social strength metric SS outperforms the other three metrics in both 2- and 3-hop social distance. Two ways in which SS differs from the other three metrics is that it considers the following attributes that the other metrics do not:

- SS considers the strength of direct ties, that is, the weights on the edges. More importantly, SS includes edge weight even in a longer social distance that is seldom taken into account by other metrics.
- SS considers that relationships are inherently asymmetric. Specifically, SS uses in its calculation the fact that for a direct tie A-B,

Table 2
Results of link prediction between pairs of n-hop distant users. Adamic-Adar: AA, Jaccard: J, Social Strength: SS. Only the SS and Katz metrics are applicable to $n = 3$.

Classifier	n	Network	Metric	Precision	Recall	F-measure	AUC
Decision tree	2	TF2	SS	0.746 ± 0.01	0.741 ± 0.08	0.744 ± 0.06	**0.765 ± 0.09**
			AA	0.714 ± 0.02	0.708 ± 0.03	0.710 ± 0.04	0.712 ± 0.08
			J	0.511 ± 0.01	0.514 ± 0.06	0.502 ± 0.07	0.510 ± 0.08
			Katz	0.697 ± 0.01	0.692 ± 0.01	0.691 ± 0.02	0.684 ± 0.05
		IE	SS	0.843 ± 0.01	0.840 ± 0.02	0.837 ± 0.02	**0.872 ± 0.01**
			AA	0.697 ± 0.01	0.693 ± 0.02	0.692 ± 0.06	0.695 ± 0.04
			J	0.698 ± 0.01	0.687 ± 0.04	0.682 ± 0.04	0.690 ± 0.07
			Katz	0.663 ± 0.03	0.660 ± 0.02	0.659 ± 0.02	0.659 ± 0.01
Decision tree	3	TF2	SS	0.630 ± 0.02	0.627 ± 0.01	0.624 ± 0.01	**0.644 ± 0.03**
			Katz	0.518 ± 0.07	0.621 ± 0.05	0.542 ± 0.03	0.537 ± 0.03
		IE	SS	0.659 ± 0.01	0.650 ± 0.01	0.646 ± 0.01	**0.664 ± 0.01**
			Katz	0.628 ± 0.05	0.609 ± 0.06	0.601 ± 0.06	0.623 ± 0.07

Table 3
Results of link prediction between pairs of n-hop distant users. Symmetric Social Strength: SymSS, Unweighted Social Strength: UWSS. Results of SS are copied from Table 2 for comparison convenience.

Classifier	n	Network	Metric	Precision	Recall	F-measure	AUC
Decision tree	2	TF2	SS	0.746 ± 0.01	0.741 ± 0.08	0.744 ± 0.06	**0.765 ± 0.09**
			UWSS	0.703 ± 0.04	0.702 ± 0.04	0.702 ± 0.05	0.739 ± 0.06
			SymSS	0.687 ± 0.02	0.681 ± 0.03	0.679 ± 0.04	0.676 ± 0.08
		IE	SS	0.843 ± 0.01	0.840 ± 0.02	0.837 ± 0.02	**0.872 ± 0.01**
			UWSS	0.686 ± 0.01	0.681 ± 0.06	0.678 ± 0.04	0.703 ± 0.01
			SymSS	0.666 ± 0.01	0.664 ± 0.02	0.664 ± 0.06	0.668 ± 0.04
Decision tree	3	TF2	SS	0.630 ± 0.02	0.627 ± 0.01	0.624 ± 0.01	**0.644 ± 0.03**
			UWSS	0.638 ± 0.03	0.611 ± 0.03	0.561 ± 0.04	0.625 ± 0.03
			SymSS	0.609 ± 0.05	0.580 ± 0.05	0.550 ± 0.08	0.585 ± 0.06
		IE	SS	0.659 ± 0.01	0.650 ± 0.01	0.646 ± 0.01	**0.664 ± 0.01**
			UWSS	0.671 ± 0.01	0.650 ± 0.03	0.638 ± 0.04	0.641 ± 0.04
			SymSS	0.640 ± 0.05	0.634 ± 0.06	0.634 ± 0.06	0.637 ± 0.07

where A has a degree d_A and B has a degree d_B, with $d_B <<d_A$, A is more important to B than B may be for A. This asymmetry translates easily to larger distances as well.

Next we investigate empirically how each of these two attributes affects the accuracy in quantifying the indirect tie strength between two nodes.

4.3.1. *Experiments*

To isolate the effects of edge weights and asymmetry, we introduce two modifications to the social strength metric *SS* in the following ways:

- We consider an unweighted social graph for the definition of *SS* in Eq. 1. Consequently, all edge weights are set to 1. We refer to this modified, *unweighted* social strength measure as *UWSS*.
- We consider a further modification of *USS* in which the asymmetry is removed. Note that asymmetry in the definition of the social strength metric is caused by normalization, that is, a user's interactions are normalized to the total number of interactions that the user has with other individuals (Eq. 2). To isolate the effect of asymmetry, we define the metric *SymSS* based on *SS*, in which $NW = 1$.

We repeat the previous experiments with the simplified metrics UWSS and SymSS using the same experimental setup as described earlier in Section 4.2.1. However, instead of SS, we use the scores calculated from UWSS and SymSS as features in the prediction task.

4.3.2. *Results*

The results shown in Table 3 allow us to make the following two observations. First, compared to the prediction results of SS in Table 2, the prediction performance of unweighted social strength's (UWSS) declines throughout both datasets and in both 2- and 3-hop social distance, and the reduction in performance reaches 20.4% (2 hops link prediction of IE). This fact confirms that edge weight, as a good social relationship proxy, is useful for evaluating social ties more accurately.

Second, for both datasets and social distances, predictions based on asymmetric social relationships (UWSS) achieve better performance than SymSS that simply treats social relationships equally. Consequently, using local graph topology information, as captured in UWSS, improves the estimation of indirect tie strength.

To conclude, compared to the social strength metric (SS) that uses the edge weights and in an asymmetrical, normalized way, the decreased performance of the unweighted version (UWSS) and the symmetrical version (SymSS) verifies that edge weights and asymmetry should be considered in a tie strength measurement. These are the very merits of our social strength metric and they lead to increased accuracy of link prediction.

5. Predicting information diffusion paths

Information diffusion is a fundamental process in social networks and has been extensively studied in the past (e.g., [36–40]). In fact, some studies have shown that the evolution of a network is affected by the diffusion of information in the network [39] and vice versa [38]. Our results from the previous study showed that indirect ties affect the process of network evolution [41]. In this section we go a step further and investigate if the strength of indirect ties can predict diffusion paths between distant nodes in the graph. That is, departing from the step-wise diffusion processes examined in the past, and given that a user received a piece of information at time t, can we predict which other users will receive this information at $t + n$ $(n \geq 2)$? Predictions over such longer intervals could help OSN providers customize strategies for preventing or accelerating information spreading. For example, to contain rumors, OSN providers could block related messages sent to the susceptible users several time steps before the rumor arrives, or disseminate official anti-rumor messages in advance. Similarly, marketers could accelerate their advertisements spreading in the network by discovering who will be the next susceptible to infection. This n-hop path prediction can supply more time for decision makers to contain harmful disseminations, and to choose users who are pivotal in information spreading for targeted advertisements.

This section describes our experiments of applying the social strength metric to information diffusion path prediction.

5.1. The Higgs-Twitter dataset

The Higgs dataset includes a 7-day scientific rumor diffusion process on Twitter in 2012 [42]. The announcement of the discovery of Higgs boson on Twitter triggered a large-scale information propagation about this topic. The dataset was collected between 1st and 7th July 2012, including four diffusion periods (before, during and after the announcement) of the event. Only the messages posted on Twitter about this discovery containing keywords or hashtags related to the Higgs event are considered as spread information. As retweets are highly relevant for the viral propagation of information [43], we use them to capture the process of diffusion. Additionally, our social strength is based on social ties thus follower-followee (FF) relationships are necessary to estimate the strength of social ties between users. We combine this relationship information with retweeting information and construct a follower-followee-retweet (FF-RT) network, which is the intersection of follower-followee and retweet networks. We apply our social strength on the FF-RT network to predict information diffusion paths. The statistics of all three networks are shown in Table 4.

5.1.1. Predicting diffusion paths via social strength

The strength of an indirect tie decreases with the length of the shortest path between the two individuals and people can be influenced up to 3 hops. Thus, we set our experiments up to 3 hops. A single node is chosen as the original source of information at t_0. We then predict the nodes that will accept the information at t_n ($n = 2$ or 3) with the knowledge from t_0.

In information diffusion, users are classified into three categories: seeds, information spreaders and non-spreaders. Thus, we divided all users in the Higgs-Twitter dataset into seeds who are the source of diffusion and never retweeted other users' messages during the diffusion process; information spreaders are users who retweeted other users' messages after exposure to them; non-spreaders are users who exposed to the information but did not retweet messages. Based on the classification of users, we extracted 2-hop and 3-hop diffusion paths from the dataset as ground truth, i.e., each directed 2(3)-hop diffusion path begins from a seed and ends with a spreader. Note that on the diffusion paths all users are spreaders. To do a binary classification (i.e., diffused or not), we also extracted 2- and 3-hop non-diffusion paths, which begin from seeds but end with non-spreaders and not all users on the path are spreaders. In our problem, we try to use indirect ties to predict end recipients status, without knowing the status of intermediary nodes. Thus, intermediary nodes could have two statuses (spread the information or not), i.e., nodes between seeds and end users could be either spreaders or non-spreaders. Once we extracted the ground truth, we then use social strength to predict whether a path is a diffusion path or a non-diffusion one. We calculate social strength values between seed and its n-hop nodes, then use the value as a feature for prediction.

In the Higgs-Twitter dataset, 22,262 users are seeds, and 204,709,636 are 2-hop non-diffusion paths while only 10,619 are diffusion paths. For 3-hop paths, 290,553,709 are non-diffusion paths and 215,445 are diffusion paths. To handle this imbalance, we under sample non-diffusion paths to make both types of paths appear with the same frequency.

5.2. Results

We compare the prediction results with the ground truth obtained from the Higgs event diffusion process to verify the effectiveness of the social strength in predicting diffusion paths. We evaluate our method using accuracy, sensitivity and specificity [44]. To better demonstrate social strength's effective power on inferring diffusion processes, we compare the social strength prediction performance with the three other metrics (Jaccard Coefficient, Adamic-Adar and Katz) introduced in Section 4.1. Note that in the Higgs-Twitter dataset, users' retweet behavior happened during the diffusion process, and users' interaction information before diffusion is not available, which means all graphs are unweighted graphs. Therefore, we use the unweighted social strength (UWSS) metric introduced in Section 4.3 instead of the weighted one. Table 5 presents the prediction results in a 2- and 3-hop social distance, respectively. As both Jaccard Coefficient and Adamic-Adar metrics are restricted to predict within 2 hops, thus, only social strength and Katz results appear for 3 hops. We see that for both 2- and 3-hop path predictions, overall the accuracies of the social strength metric are higher than the other three metrics' in most of the scenarios, reaching a maximum of 0.753 with social strength metric in the 2-hop path prediction. The only exception occurs with 3-hop paths prediction where Katz shows minor advantage than social strength. Although 3-hop predictions show decreased accuracy compared to 2-hop results, they remain above 0.50.

From all these results, we conclude that indirect ties have potential in controlling the flow of information in the network that should not be ignored. More importantly, our social strength, as an indirect tie measurement, is useful to predict who will be the spreader, or along which paths information propagates, at least 2–3 steps before a susceptible node is even in contact with a spreading node.

6. Using social strength in friend-to-friend storage systems

F2F systems are distributed systems where social incentives encourage users to provide resources from their local machines to their friends. For example, Tribler [45] is a friend-based P2P file sharing system, which relies on friends' similar taste in content to encourage altruistic behavior; Turtle [46] leverages users' pre-existent trust relationships to supply safe sharing of sensitive data.

Although a promising alternative to cloud-based data backup, F2F storage systems were shown to suffer from two significant limitations. First, users with a small set of friends are penalized by lack of available storage for their needs, resulting in low resource utilization [4]. Second, friends are typically in close geographical proximity, and thus their online times are synchronized, leading to high unavailability to their friends' data [14]. These concerns can intuitively be

Table 4
The statistics of networks for diffusion paths predictions. ACC: average clustering coefficient and OT: observation time.

Networks	Nodes	Edges	ACC	Diameter	OT
Follower-followee (FF)	456,631	14,855,842	0.1887	9	N/A
Retweet (RT)	256,491	328,132	0.0156	19	7 days
Follower-followee-retweet (FF-RT)	254,872	320,467	0.0155	19	N/A

Table 5
Results of link prediction between pairs of n-hop distant users. Adamic-Adar: AA, Jaccard: J, Unweighted Social Strength: UWSS. Only the UWSS and Katz metrics are applicable to $n = 3$.

Network	n	Classifier	Metric	Precision	Recall	F-measure	AUC
Higgs-Twitter	2	Decision tree	UWSS	0.692 ± 0.001	0.651 ± 0.003	0.631 ± 0.004	**0.653 ± 0.006**
			AA	0.616 ± 0.005	0.585 ± 0.008	0.555 ± 0.005	0.642 ± 0.004
			J	0.621 ± 0.001	0.621 ± 0.001	0.620 ± 0.004	0.621 ± 0.002
			Katz	0.500 ± 0.000	0.500 ± 0.000	0.405 ± 0.000	0.500 ± 0.000
		Logistic regression	UWSS	0.695 ± 0.002	0.674 ± 0.002	0.664 ± 0.002	**0.753 ± 0.001**
			AA	0.521 ± 0.001	0.515 ± 0.001	0.480 ± 0.001	0.510 ± 0.002
			J	0.619 ± 0.001	0.611 ± 0.001	0.604 ± 0.001	0.678 ± 0.001
			Katz	0.726 ± 0.004	0.502 ± 0.000	0.339 ± 0.000	0.502 ± 0.000
	3	Decision tree	UWSS	0.695 ± 0.037	0.566 ± 0.026	0.471 ± 0.049	**0.578 ± 0.029**
			Katz	0.500 ± 0.001	0.50 ± 0.001	0.405 ± 0.001	0.500 ± 0.001
		Logistic regression	UWSS	0.637 ± 0.126	0.609 ± 0.103	0.581 ± 0.114	0.623 ± 0.120
			Katz	0.606 ± 0.073	0.589 ± 0.070	0.552 ± 0.102	**0.628 ± 0.063**

addressed by leveraging social strength (SS_n) to expand the set of resources while still using a measure of social incentives.

In this section, we verify whether our social strength metric can improve the service performance in F2F storage systems. To maintain a meaningful value of social incentives, we restrict our evaluations to $n = 2$ and $n = 3$. Our objectives are:

- To understand if SS_n expands the size of candidate sets.
- To evaluate the benefits of using SS_n to improve data availability in F2F systems.

6.1. Social strength expands users' friendsets

In the following, we experimentally show how social strength can be used for expanding users' friendsets in the system.

6.1.1. Friendset expansion algorithm

Some socially aware systems have explored indirect ties among users in the design of their systems. Some previous work directly involves all of a user's friends-of-friends (or even longer distant relationships) [45]. This naive friend sets expansion scheme could enlarge many users' friend sets, specifically in networks where most of their nodes' shortest paths are larger than the length of the expansion. However, not every 2- or 3-hop distant friend has enough incentives, for example, for storing data, hosting computation tasks, or routing messages. Therefore, instead of directly involving all of a user's indirect ties within some radius, we use the quantitative power of the social strength metric, SS_n, to select socially "close" distant nodes, that is, indirect connections with comparable social strength with the user's direct (1-hop) friends. However, even a user's indirect contacts could have no willingness to share their resource. To consider this uncertainty of resource sharing among indirect ties, we introduce a degree of uncertainty u ($u \in [0, 1]$) into the estimation of trust based on the expansion algorithm in [47].

The expansion algorithm follows three steps:

- For each user i, find the weakest direct social contact p such that $NW(i, p) = \min_{j \in Neigh(i)} [NW(i, j)]$. Let this minimum normalized weight be referred to as θ_i.
- For each m of i's n-hop friends, if $SS_n(i, m) \geq \theta_i$, the user m is inserted in the candidate peer set of i. Intuitively, this ensures that the social strength between i and m, located at distance n in the social graph, is at least as strong as i's weakest direct tie.
- For each users' n-hop friend peer set, only 1-u of this set's peers are randomly selected as trust candidates for resource sharing.

We note that the algorithm expands each candidate set using a user-specific, thus local, threshold. Such local thresholds are needed in the distributed setting of a F2F system.

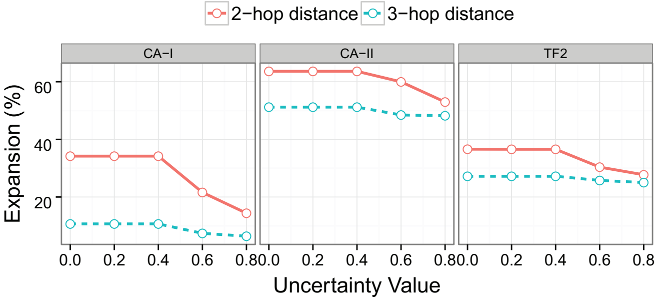

Fig. 3. Candidate set expansion via SS_2 (only 2-hop friends) and SS_3 (only 3-hop friends): percentage of expanded users with different values of uncertainty in CA-I, CA-II and TF2.

6.2. Datasets

The online game interacting friends network, TF2 (introduced in Section 3) also supplies each player's online/offline status that can be used for the data availability experiments presented later, in Section 6.4. However, the face-to-face contact network of Infectious Exhibition (IE) (used in Section 3) is an ephemeral offline social network, which does not include any users' online activities. Thus, the IE network is not suitable for our later experiments, and thus we did not use it for the evaluation of the friendset expansion. Instead, we use two co-authorship networks CA-I and CA-II. Nodes in this graph represent authors and are labeled with the author's affiliation. We map each author's affiliation information to a timezone, which can be further used in simulating users' online/offline behaviors in Section 6.4.1. To expand users' peer sets with uncertainty, we tested a wide range of uncertainty to cover possible cases in trust estimation, i.e., $u = 0.1$–0.8.

6.3. Expansion results

Since the most intuitive advantage of our mechanism is an increase in the number of storage candidates, we begin by evaluating *how much* the candidate set is expanded. We thus implemented $SS_n(i, m)$ presented in Section 2 and report the size of the candidate set selected based on the expansion algorithm presented in Section 6.1.1 on the three networks described in Section 6.2.

Fig. 3 shows how candidate sets are expanded with 2- and 3-hop social distance respectively in each of our three networks. For 2-hop expansion without uncertainty, 63.62% users in CA-II and 36.6% of players in TF2 expanded their candidate sets. Even in the sparse CA-I, 34.19% users augmented their friend sets. After adding uncertainties in the friendset expansion algorithm, the percentage of expanded users decreases. But this only happens when the degree of uncertainty (u) is larger than 0.6, that is more than 60% of selected peers refuse to share their resource. The degree of uncertainty barely influences expansions when the uncertainty value is smaller than 0.6.

When considering the expansion (no uncertainty is considered) brought in by 3-hop distant nodes p who satisfy the requirement that $SS_3(i, p) \geq \theta(i)$ the expansion is still taking place in all three networks: even in the sparse network CA-I, 10.6% users augment their friendsets and about 1% users have expanded their candidates with more than five friends. The denser network CA-II has more than 50% of users expanding their candidate sets, and TF2 has 27.2% (with the number of expanded 3-hop friends being 1,032). With the increase of uncertainty, the sparse CA-I has more percentage of users reduce their expanded peers, from 10.67% with zero uncertainty to 6.4% with uncertainty value of 0.8, compared to denser CA-II and TF2. However, 3 hops' expansion declines slower than 2 hops. This is because a user could be expanded with more candidates in 3 hops than 2 hops then even a high degree of uncertainty is added, the user still has at least one peer to expand his friendset.

All in all, as expected, 3-hop augmentation is not as strong as 2 hops' since as the social distance increases, the social strength weakens. Yet a number of users can still recruit more peers when increasing the social distance. In addition, if users have a large number of peers for resource sharing, a small (or median) degree of uncertainty seldom affects friendset expansions. Thus, using social strength for recruiting peers indirectly connected in the social graph augments users' peer-sets and potentially solves problems caused by the limited number of friends in F2F systems.

6.4. *Expanded friendsets improve data availability*

Expanding the candidate set is a necessary but insufficient solution for improving the performance of F2F systems. In particular, as shown in the context of F2F storage systems, F2F service availability depends on user online activity patterns [14,48].

In this section we show that a larger resource candidate set can significantly improve data availability in F2F systems. We stress that we do not propose a cohesive mechanism that improves the performance of F2F systems. Instead, we focus on exploring the potential of using social strength (via the expansion algorithm) in F2F solutions. Thus, the following sections show that data availability under a previously proposed replica allocation strategy increases significantly compared with "traditional" 1-hop F2F. In the following section, our augmented candidate (friends) sets refer to users that have expanded their friend sets with our expansion algorithm up to 3-hop social distance.

6.4.1. *Online presence behavior*

We simulate users' online presence and data placement to estimate file availability in F2F storage systems with service candidate sets augmented by social strength. To estimate peer availability, we augment each network with online presence empirically deduced from real traces. For CA-I and CA-II, we fit a distribution to online presence information extracted from empirical Skype traces presented in [14]. The distribution was applied to each author by shifting it to match the timezone of his or her affiliation. As seen in Fig. 4, which plots the percentage of users online per hour of the day, at least 25% of nodes are online at any given time, with the peak and valley occurring at about 1:00 AM and noon, respectively. For the TF2 network, we use one month of recorded playing times. We plot the corresponding aggregate distribution in Fig. 5, which shows each week's online presence per hour for May 2011. The distribution shows clear diurnal and domain-specific activity patterns. As noted in [22], gaming is not an activity conducive to multi-tasking. Therefore, we see an elevated level of presence on weekends and during non-working hours. Although peak presence occurs consistently in the early morning with more than 20% of users online, there are almost no users online at noon.

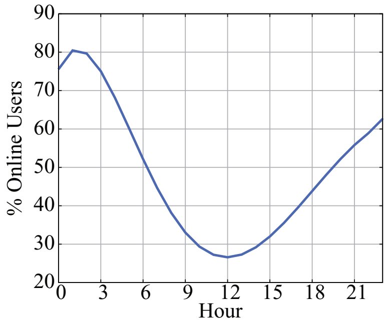

Fig. 4. Online behavior of nodes in empirical traces of Skype.

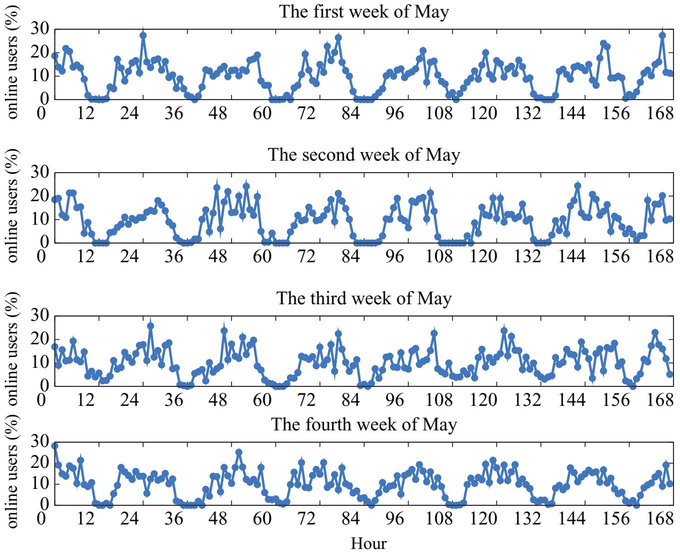

Fig. 5. Online behavior of players per hour of the week in May for TF2.

To determine whether the social strength selection mechanism improves the availability of storage resources, we measure the percentage of a node's selected candidates available throughout the day, by binning online presence into 1 hour time slots. We also map each user's affiliation to a timezone, then match the timezone to an hour of a day. If a user is online at some point during a time slot, we mark him as available for that time slot. For CA-I and CA-II networks, in each time slot, we randomly select users to be online but keep the same percentage of online users from the Skype trace. We repeat this random sampling process for multiple iterations to obtain stable results. Methods that store files in a distributed fashion such as erasure codes require k storage sites to be available for retrieving a file [49]. Thus, we also vary the number of friends necessary for a node's storage needs to be met under such storage schemes. We then measure the fraction of nodes who have enough candidates online to meet their needs when selected by either the pure F2F approach or the social strength mechanism.

6.4.2. *Data placement*

Replicating data across all friends allows a user to get maximum achievable data coverage but results in high costs for storing and transferring data to multiple copies, in particular for users with a large number of friends. So we adopt the greedy heuristic data placement algorithm proposed in [48] to backup files with a subset of friends who can cover the maximum online time. In this heuristic, to get maximum possible time slots coverage (e.g., 24 hours), users

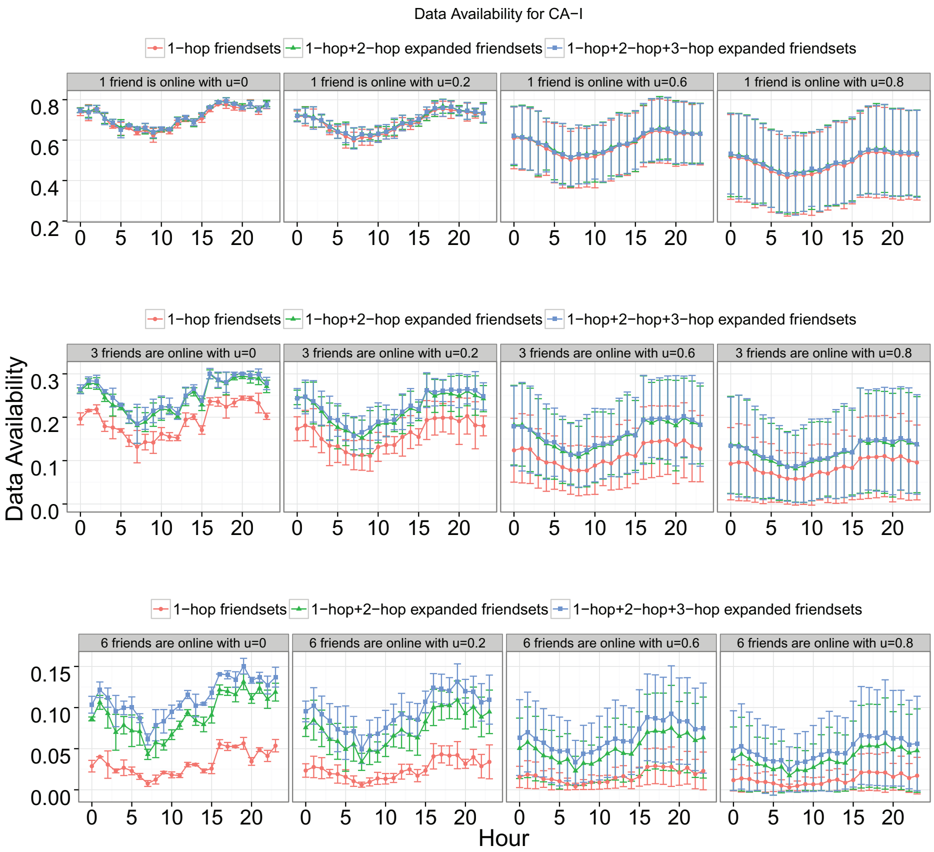

Fig. 6. Average fraction of available candidates per hour for CA-I, and u is the parameter for uncertainty. For brevity, only results of $u = 0.2$ are presented since $u = 0.4$ performs similarly with 0.2.

first pick a set of friends who are able to cover at least one unique time slot that other friends cannot cover. If this set of friends cannot cover all the time slots, then select other friends to cover the remaining uncovered time slots and keep doing this until all the time slots are covered or no friends can cover the uncovered time slots.

6.4.3. *Data availability*

Some methods store files in a distributed fashion such as erasure codes that require k storage sites to be available for retrieving a file [21]. We vary the number of friends necessary for a node's storage needs to be met under such storage schemes. We measure the fraction of nodes who have enough candidates online to meet their needs when selected by either the pure F2F approach or the social strength mechanism. We compare three scenarios: 1) storage candidates selected only from direct social contacts; 2) storage candidates selected from the SS_n-based expanded candidate set,with $n = 2$ and 3) $n = 3$. Figs. 6, 7 and 8 plot the average fraction of users whose storage needs are met with the requirement that at least $k \in \{1, 3, 6\}$ candidates are online at a given time for the co-authorship networks and TF2, respectively. Error bars represent the 95% confidence interval on average. Three scenarios are compared: 1) storage candidates selected only from direct social contacts, storage candidates selected from the SS_n-based expanded candidate set, with 2) $n = 2$ and 3) $n = 3$. A range of different degrees of uncertainty (u) is considered, i.e., $u = 0.1 – 0.8$.

Using the expanded candidate set results in higher service availability. In particular, when 6 friends are needed to cooperate on completing a storage task, about 4 times higher data availability can be reached in CA-I, 1.6 times higher in CA-II and 6.5 times higher in TF2. Further, the social strength mechanism does not degrade as quickly as the 1-hop selection when increasing the number of friends that are required to be online simultaneously. We also see that for sparse networks like CA-I, social strength over larger distance n improves data availability, especially when larger number of friends are required to be online simultaneously.

Finally, CA-II shows higher levels of availability than CA-I under the same conditions. This is likely because CA-II has more users with larger expanded candidate sets under the social strength mechanism than CA-I (Fig. 1). Moreover, we note that CA-I shows better performance than TF2 under the same requirements. In the scenario that requires at least one friend online, 73% of users in CA-I have candidates available at midnight, compared to only 20% of TF2 users. One explanation could be the limited number of concurrent players the gaming server supports (at most 32 simultaneous players). Another explanation is that CA-I users are spread out over multiple timezones, while most of the TF2 users are geographically close to the server to minimize latency, and thus are time synchronized in their gaming patterns.

For scenarios with uncertainty, when uncertainty is 0.2 and 0.4, data availability experiences almost no change in all three datasets. Even when the uncertainty increases to 0.6, less than 15% of data availability is reduced. With the uncertainty degree increased to 0.8, data availability is reduced at most by 23%.

To conclude, using datasets from co-authorship networks and a video gaming community, we show that the social strength-based algorithm more than doubles the set of storage candidates potentially motivated by social incentives, and increases data availability by up to three times compared to the pure F2F approach.

7. Related work

In sociology, two theories are closely related to the properties of social ties. First, the theory of *homophily* [26] postulates that people

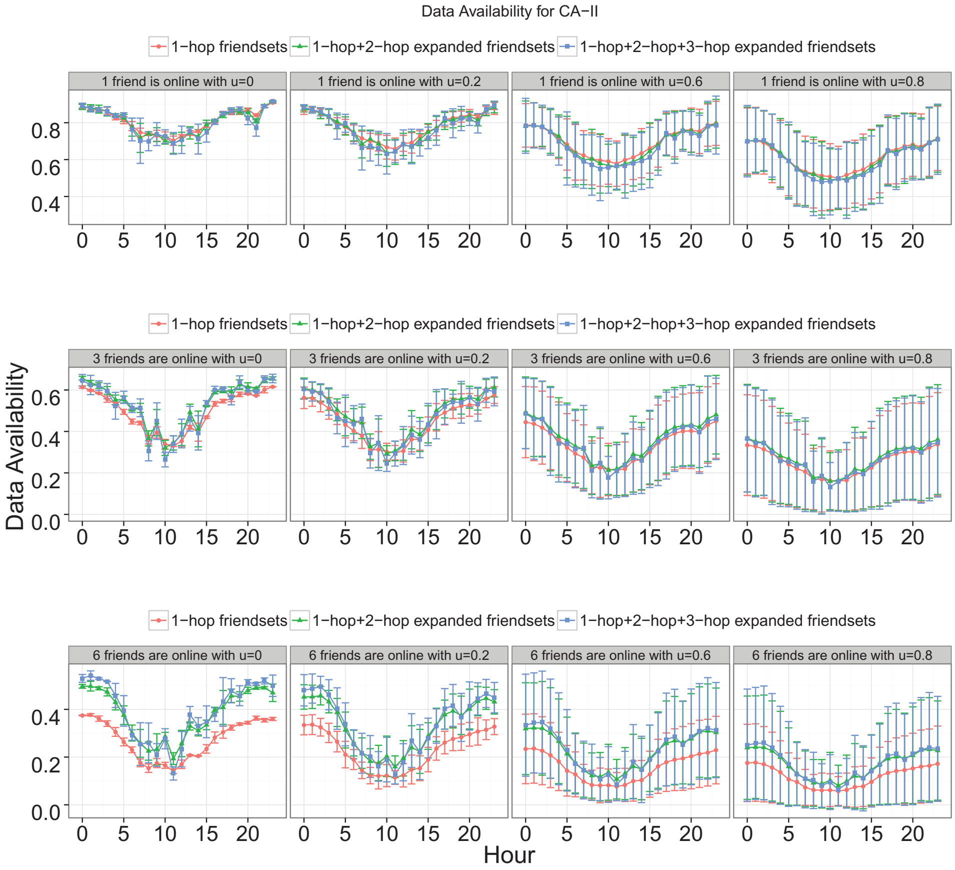

Fig. 7. Average fraction of available candidates per hour for CA-II, and u is the parameter for uncertainty. For brevity, only results of $u = 0.2$ are presented since $u = 0.4$ performs similarly with 0.2.

tend to form ties with others who have similar characteristics. Moreover, a stronger relationship implies greater similarity [8]. Second, the principle of *triadic closure* [50] states that two users with a common friend are likely to become friends in the near future. The triadic closure has been demonstrated as a fundamental principle for social network dynamics. For example, Kossinets and Watts [51] showed how it amplifies homophily patterns by studying the triadic closure in e-mail relations among college student. Kleinbaum [52] found that persons with a typical careers in a large firm tend to lack triadic closure in their email communication network and so have their brokerage opportunities enhanced.

Since Granovetter [8] introduced the notion of strength of ties in social networks, there have been many studies on tie strength measurement. Gilbert and Karahalios [6] modeled tie strength as a combination of social dimensions such as intensity, intimacy, duration, and structure. Crandall et al. [53] investigated the existence of social ties between people from co-occurrence in time and space on Flickr and discovered that even a small number of co-occurrences indicate a high probability of an existing tie between two users. Likewise, Kahanda and Neville [5] developed a supervised learning predictor that classifies a link in OSNs as either a weak or strong tie via features from user profiles, graph topology, transactional connectivity and network-transactional connectivity features.

However, these methods either need extra information (e.g., users' profiles, the message content or users' geo-locations) or adopt complex models that cannot be implemented in a decentralized fashion. More importantly, most previous methodologies simply treat users' relationships symmetrically. Without asymmetric discrimination, it is difficult to accurately capture the strength of social ties [21].

Social networks as a channel for people to share information have been studied extensively in the context of information diffusion, especially the role of tie strength in diffusion. Aral et at. [54] pointed out that whether or not information is delivered through a tie depends on the tradeoff between structure diversity and "bandwidth" (interaction frequency). Grabowicz [55] empirically observed that intermediary social ties are a vital component in information diffusion of online social networks. Bakshy et al. [56] compared the role of strong and weak ties in information propagation and found that weak ties dominate the propagation process instead of strong ties that were originally believed. Levin et al. [57] surveyed three companies to prove that weak ties, providing access to non-redundant information, are more useful for information diffusion. Although most of these studies concentrated on directly connected social ties, they provide foundations for our study and motivate us to investigate the relationship between indirect ties and information diffusion.

This work provides an indirect tie metric that only needs graph topology information and can be implemented in a decentralized fashion. Most importantly, this work contributes the validation of the social metric and demonstrates its value via proof-of-concepts applications that use it.

8. Summary and discussions

In this paper, we introduced a social strength metric to measure the strength of indirect social ties by considering both the intensity of interactions and the number of connected paths. We showed that our metric is effective in predicting links formation (can achieve 0.881 prediction accuracy), indicating that it is an accurate quantification of the intensity of an indirect social relationship.

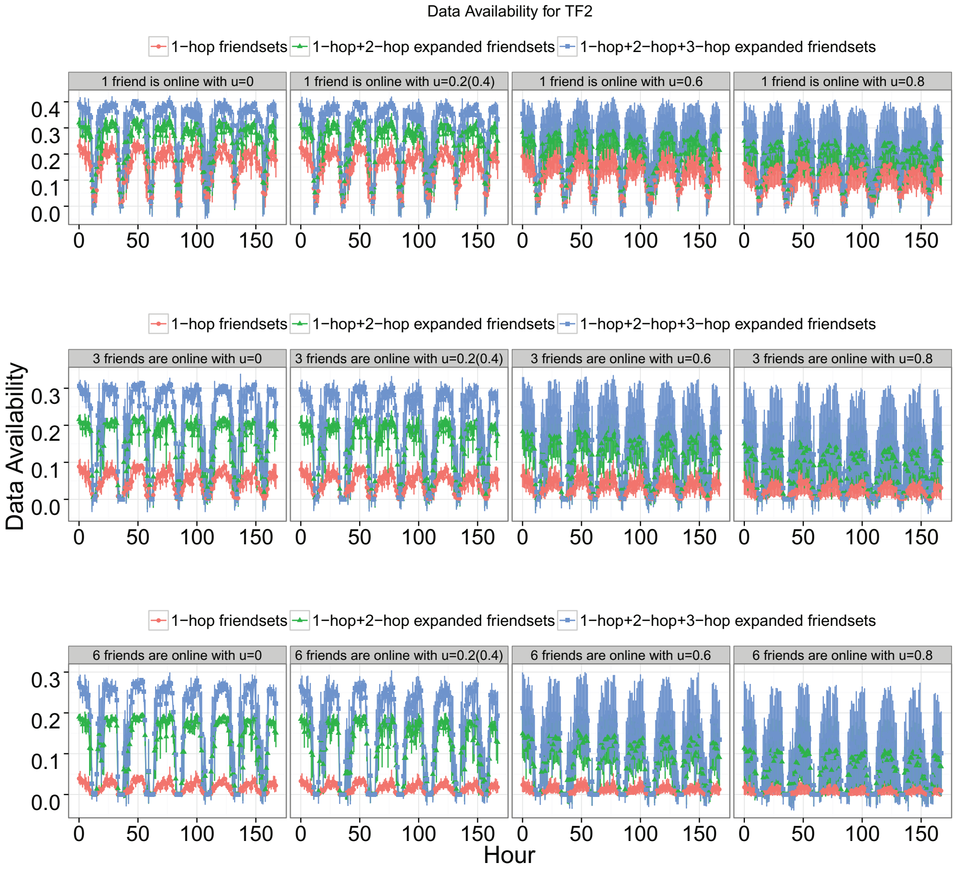

Fig. 8. Average fraction of available candidates per hour of the week for TF2, and *u* is the parameter for uncertainty.

Further, we proved our proposed metric's applicability to two socially informed applications: predicting information diffusion in a social graph and friend-to-friend storage sharing systems. Based on empirical data, our experimental evaluations demonstrate that using the social strength metric is beneficial in both cases. First, social strength accurately predicts information diffusion paths at least 2 steps ahead, which enables intervention mechanisms for rumor squelching and targeted information injection. Second, for the average user in the social graph, it helps identify indirectly connected peers with whom the user has a significant social strength that could act as social incentive in a resource sharing environment, thus significantly increasing the pool of resources available to the user. Third, because indirect ties diversify the pool of users (in this case, by covering more time zones), resource availability increases significantly.

A variety of socially aware applications can benefit from the social strength metric. For example, link prediction based on social strength could discover more potentially useful contacts and improve link recommendation accuracy. Automatically setting default privacy controls based on social strength is likely to be more accurate than using graph distance alone. Employing social strength in graph partitioning will have the benefits of relying on local computation, thus allowing for more decentralized and scalable algorithms. Finally, in decentralized OSNs, users' augmented social strength-based friendsets could provide a more efficient and privacy-guaranteed technique to propagate updates in the presence of churn.

This work is a first step in understanding the value of and the methodology for quantifying the strength of indirect social ties. In addition to exploring the applicability space, there are aspects related to privacy and security that need to be understood. Intuitively, because of the local exploration of one's social neighborhood for computing social strength, the risks are contained, especially compared to approaches that require the global graph. However, a formal study of this topic is required for building a practical framework that enables the implementation and adoption of the social strength metric for indirect ties.

Acknowledgment

This research was supported by the U.S. National Science Foundation under grant no. CNS 0952420 and by the MULTISENSOR project partially funded by the European Commission under contract number FP7-610411.

References

[1] G.C. Homans, Social behavior: its elementary forms, Harcourt, Brace, 1961.
[2] Z. Li, H. Shen, Soap: a social network aided personalized and effective spam filter to clean your e-mail box, in: INFOCOM, Proceedings IEEE, 2011, pp. 1835–1843.
[3] C. Basu, H. Hirsh, W. Cohen, Recommendation as classification: Using social and content-based information in recommendation, in: AAAI/IAAI, 1998, pp. 714–720.
[4] J. Li, F. Dabek, F2F: reliable storage in open networks, in: Proceedings of the 4th International Workshop on Peer-to-Peer Systems (IPTPS), 2006, pp. 1–10.
[5] I. Kahanda, J. Neville, Using transactional information to predict link strength in online social networks, in: ICWSM, 2009, pp. 1–10.
[6] E. Gilbert, K. Karahalios, Predicting tie strength with social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2009, pp. 211–220.
[7] R.S. Burt, Social contagion and innovation: cohesion versus structural equivalence, American j. Soc. (1987) 1287–1335.
[8] M.S. Granovetter, The strength of weak ties, American J. Soc. 78 (6) (1973).
[9] M. Granovetter, Getting a job: a study of contacts and careers, University of Chicago Press, 1995.
[10] P. Anderson, N. Koutellis, J. Finnis, A. Iamnitchi, On managing social data for enabling socially-aware applications and services, in: 3rd Workshop in Social Network Systems, 2010, pp. 1–10.
[11] N. Kourtellis, On the Design of Socially-Aware Distributed Systems, University of South Florida, 2012 Ph.D. thesis.

[12] B. Wellman, Structural analysis: From method and metaphor to theory and substance, Social struct.: a network approach. (1988) 19–61.

[13] N.E. Friedkin, Horizons of observability and limits of informal control in organizations, Social Forces 62 (6) (1983) 54–77.

[14] G.T. Raúl, M. Sánchez Artigas, P. García López, Analysis of Data Availability in F2F Storage Systems: When Correlations Matter, in: Peer-to-Peer Computing, 2012, pp. 225–236.

[15] G.C. Homans, The human group, 7, Routledge, 2013.

[16] C. Wilson, B. Boe, A. Sala, K.P.N. Puttaswamy, B.Y. Zhao, User interactions in social networks and their implications, in: Proceedings of the 4th ACM European conference on Computer systems, 2009, pp. 205–218.

[17] N.A. Christakis, J.H. Fowler, Connected: the surprising power of our social networks and how they shape our lives, Hachette Digital, Inc., 2009.

[18] N.A. Christakis, J.H. Fowler, The spread of obesity in a large social network over 32 years, New England j. med. 357 (4) (2007) 370–379.

[19] J.H. Fowler, N.A. Christakis, D. Roux, Dynamic spread of happiness in a large social network: longitudinal analysis of the framingham heart study social network, BMJ: British med. j. (2009) 23–27.

[20] L. Pappalardo, G. Rossetti, D. Pedreschi, How well do we know each other? detecting tie strength in multidimensional soical networks, in: In ASONAM, 2012, pp. 1040–1045.

[21] D.J. Brass, K.D. Butterfield, B.C. Skaggs, Relationships and unethical behavior: a social network perspective, Acad. Manag. Rev. 23 (1) (1998) 14–31.

[22] J. Blackburn, A. Iamnitchi, Relationships under the microscope with interaction-backed social networks, in: 1st International Conference on Internet Science, 2013, p. 199.

[23] J. Blackburn, N. Kourtellis, J. Skvoretz, M. Ripeanu, A. Iamnitchi, Cheating in online games: a social network perspective, ACM Trans. Int. Tech. (TOIT) 13 (3) (2014) 9.

[24] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, W.V. den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks, J. Theor. Biol. 271 (1) (2011) 166–180.

[25] J. Tang, J. Sun, C. Wang, Z. Yang, Social influence analysis in large-scale networks, in: International Conference on Knowledge Discovery and Data Mining (KDD), 2009, pp. 807–816.

[26] M. McPherson, L. Smith-Lovin, J. Cook, Birds of a feather: homophily in social networks, Ann. rev. soc. (2001) 415–444.

[27] L. Lü, T. Zhou, Link prediction in complex networks: a survey, Physica A: Stat. Mech. Appl. 390 (6) (2011) 1150–1170.

[28] R. Xiang, J. Neville, M. Rogati, Modeling relationship strength in online social networks, in: 19th International Conference on World Wide Web, Raleigh, NC, USA, 2010, pp. 981–990.

[29] G. Salton, M.J. McGill, Introduction to modern information retrieval, New York: McGraw-Hill (1983).

[30] L. Adamic, E. Adar, Friends and neighbors on the web, Soc. net. 25 (3) (2003) 211—230.

[31] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.

[32] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. American soc. inf. sci. tech. 58 (7) (2007) 1019–1031.

[33] M. Kubat, S. Matwin, et al., Addressing the curse of imbalanced training sets: one-sided selection, in: ICML, 97, 1997, pp. 179–186.

[34] N.V. Chawla, K.W. Bowyer, H.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[35] I.H. Witten, Eibe, Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2005.

[36] M. Yildiz, A. Scaglione, A. Ozdaglar, Asymmetric information diffusion via gossiping on static and dynamic networks, in: 49th IEEE Conference on Decision and Control (CDC), 2010, pp. 7467–7472.

[37] A. Guille, H. Hacid, A predictive model for the temporal dynamics of information diffusion in online social networks, in: Proceedings of the 21st International Conference Companion on World Wide Web, 2012, pp. 1145–1152.

[38] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the 21st International Conference on World Wide Web, in: WWW, 2012, pp. 519–528.

[39] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, A. Flammini, The role of information diffusion in the evolution of social networks, in: Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining, in: KDD, 2013, pp. 356–364.

[40] A. Guille, H. Hacid, C. Favre, D.A. Zighed, Information diffusion in online social networks: A survey, SIGMOD Rec. 42 (2) (2013) 17–28.

[41] X. Zuo, J. Blackburn, N. Kourtellis, J. Skvoretz, A. Iamnitchi, The influence of indirect ties on social network dynamics, in: To be appear in the 6th International Conference on Social Informatics, 2014, pp. 1–14.

[42] M.D. Domenico, A. Lima, P. Mougel, M. Musolesi, The anatomy of a scientific rumor, Scienific reports 3 (2013).

[43] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, W. Kellerer, Outtweeting the twitterers-predicting information cascades in microblogs, in: Proceedings of the 3rd conference on Online social networks, 2010, p. 3.

[44] T. Fawcett, An introduction to ROC analysis, Pattern recognition letters 27 (8) (2006) 861–874.

[45] J.A. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D.H. Epema, M. Reinders, M.R.V. Steen, H.J. Sips, Tribler: a social-based peer-to-peer system, Con. Comput.: Prac. Exp. 20 (2) (2008) 127–138.

[46] B.C. Popescu, B. Crispo, A.S. Tanenbaum, Safe and private data sharing with turtle: friends team-up and beat the system, in: Proceedings of the 12th International conference on Security Protocols, in: SP '04, 2004, pp. 221–230.

[47] X. Zuo, J. Blackburn, N. Kourtellis, J. Skvoretz, A. Iamnitchi, The power of indirect ties in friend-to-friend storage systems, in: 14th IEEE International Conference on Peer-to-Peer Computing, 2014.

[48] R. Sharma, A. Datta, M.D. Amico, P. Michiardi, An empirical study of availability in Friend-to-Friend storage systems, in: Peer-to-Peer Computing, 2011, pp. 348–351.

[49] W.K. Lin, D.M. Chiu, Y.B. Lee, Erasure code replication revisited, in: Proceedings of the Fourth International Conference on Peer-to-Peer Computing, 2004, pp. 90–97.

[50] A. Rapoport, Spread of information through a population with socio-structural bias: I. assumption of transitivity, The bull. math. biophys. 15 (4) (1953) 523–533.

[51] G. Kossinets, D.J. Watts, Origins of homophily in an evolving social network1, American J. Soc. 115 (2) (2009) 405–450.

[52] A.M. Kleinbaum, Organizational misfits and the origins of brokerage in intrafirm networks, Admin. Sci. Quart. 57 (3) (2012) 407–452.

[53] D.J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, J. Kleinberg, Inferring social ties from geographic coincidences, Proceedings of the National Academy of Sciences 107 (52) (2010) 22436–22441.

[54] S. Aral, M.V. Alstyne, The diversity-bandwidth trade-off1, American J. Soc. 117 (1) (2011) 90–171.

[55] P.A. Grabowicz, J.J. Ramasco, E. Moro, J.M. Pujol, V.M. Eguiluz, Social features of online networks: the strength of intermediary ties in online social media, PloS one 7 (1) (2012) e29358.

[56] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the 21st international conference on World Wide We(WWW12)b, ACM, New York, NY, USA, 2012, pp. 519–528, doi:10.1145/2187836.2187907.

[57] D.Z. Levin, R. Cross, The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer, Manag. sci. 50 (11) (2004) 1477–1490.