The Influence of Indirect Ties on Social Network Dynamics

Xiang Zuo[1], Jeremy Blackburn[2], Nicolas Kourtellis[3], John Skvoretz[4], and Adriana Iamnitchi[1]

[1]Computer Science & Engineering, University of South Florida, FL, USA
[2]Telefonica Research, Barcelona, Spain
[3]Yahoo Labs, Barcelona, Spain
[4]Department of Sociology, University of South Florida, FL, USA

xiangzuo@mail.usf.edu,jeremyb@tid.es,kourtell@yahoo-inc.com,
jskvoretz@usf.edu,anda@cse.usf.edu

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 less attention. Yet in real life, we often rely on friends of our friends for recommendations (of doctors, schools, or babysitters), for introduction to a new job opportunity, and for many other occasional needs. In this work we empirically study the predictive power of indirect ties in two dynamic processes in social networks: new link formation and information diffusion. We not only verify the predictive power of indirect ties in new link formation but also show that this power is effective over longer social distance. Moreover, we show that the strength of an indirect tie positively correlates to the speed of forming a new link between the two end users of the indirect tie. Finally, we show that the strength of indirect ties can serve as a predictor for diffusion paths in social networks.

Keywords: indirect ties, social network dynamics, information diffusion

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 confirmed long standing results from sociology on large 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 [1] to recommendations [2] and peer-to-peer backup systems [3].

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 [4,5]. However, indirect social ties, defined as relationships between two individuals who have no direct

relation but are connected through a third party, carry a significantly larger potential [6].

This paper analyzes the quantifiable effects that indirect ties have on network dynamics. Its contributions are summarized as follows:

- We quantitatively confirm on real datasets several well-established sociological phenomena: triadic closure, the timing of tie formation, and the effect of triadic closure on information diffusion.
- We extend the study of the indirect ties' impact on network dynamics to a distance longer than 2 hops.
- We show that indirect ties accurately predict information diffusion paths.

The rest of paper is organized as follows. Section 2 provides the context for this work. Section 3 introduces the datasets used in this study. Section 4 shows that the strength of *indirect* ties can be used to predict the formation of direct links at longer social distance. Section 5 refines this quantification to classify an indirect tie as weak or strong, showing that the classification meets theoretical expectations of a positive correlation between the strength of a tie and the speed at which a link forms. We also show in Section 6 that pairs with a strong indirect tie end up having more interactions after link formation when compared to pairs with a weaker indirect tie. In Section 7, we examine indirect tie strength as a predictor for diffusion paths in a network. Finally, Section 8 concludes with a discussion of lessons and future work.

2 Related Work

In sociology, two theories are closely related to the properties of indirect ties. First, the theory of *homophily* [7] postulates that people tend to form ties with others who have similar characteristics. Moreover, a stronger relationship implies greater similarity [8]. Second, the principle of *triadic closure* [9] 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 [10] showed how it amplifies homophily patterns by studying the triadic closure in e-mail relations among college student. Kleinbaum [11] found that persons with atypical careers in a large firm tend to lack triadic closure in their email communication network and so have their brokerage opportunities enhanced.

Lately, large online social networks provided unprecedented opportunities to study dynamics of networks. Thus, many studies examined the evolution of groups or analyzed membership and relationship dynamics in these networks. For example, Backstrom et al. analyze how communities or groups evolve over time and how a community dies or falls apart [12]. Patil et al. use models to predict a group's stability and shrinkage over a period of time [13]. Yang and Counts examine the diffusion of information and innovations and the spread of epidemics and behaviors [14].

Compared to previous studies, we quantitatively investigate the effects of indirect ties on network dynamics, specifically on tie formation, the speed of tie

formation, and information diffusion. More importantly, we study the impact of longer indirect ties on network dynamics: while previous work focused on 2-hop indirect ties, we also show the impact of 3-hop indirect ties.

3 Datasets

In this paper we use several datasets from different domains. Our datasets are varied, 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 [15]. The dataset includes game-based interactions among players, timestamp information of each interaction, declared relationship in the associated gaming OSN, Steam Community [16], 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 three advantages. First, it provides the number of in-game interactions that can be used to quantify the strength of a social tie. Second, each interaction and friendship formation is annotated with a timestamp, which is helpful for examining the dynamics of links under formation. Third, over a pure in-game interaction network, it has the advantage of selecting the most representative social ties, as shown in [15].

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

Networks	Nodes	Edges	APL	Density	CC	A	D	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

Infectious Exhibition (IE) held at the Science Gallery in Dublin, Ireland, from April 17^{th} to July 17^{th} 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 [17]. 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 the two largest connected components of the co-authorship graph of Computer Science researchers extracted by Tang et al. [18] from ArnetMiner[1]. Nodes in these graphs represent authors, edges are weighted with the number of papers co-authored. Because the

[1] `http://arnetminer.org/`

dataset does not include time publication information, the observation window is unspecified in Table 1.

Note that IE is a smaller but much denser network than TF2, while TF2's interactions frequency is higher than IE's, as shown by the range of edge weights. We use the TF2 and IE networks to study link formation and delay as they contain timestamps of the links formed and interactions between users. We use the TF2 and CA networks to study diffusion as they are larger, sparser and based on longer lasting relationships compared to IE's ad-hoc interactions.

4 Predicting Link Formation

According to Granovetter's idea of the *forbidden triad* [8], a triad between users u, v and w in which there are strong ties between u and v and between v and w, but no tie between u and w is unlikely to exist. When it does, according to the theory of *triadic closure*, it is typically quickly closed with the formation of a tie between u and w.

In this section, we not only empirically verify the theory of triadic closure by using multiple measures of the strength of indirect ties, but we also examine this theory over paths of length 3.

4.1 Methodology

The link prediction problem asks whether two unconnected nodes will form a tie in the near future [19]. Link prediction models that use an estimation of the tie strength from graph structure [20] or interaction frequency and users' declared profiles similarities [21] have been proposed in the past.

We use a group of tie strength metrics and classifiers to quantitatively demonstrate how indirect ties can be used for inferring new links formation. Specifically, given a snapshot of a social network, we use the strength of indirect ties to infer which relationships or interactions among users are likely to occur in the near future. Because people can be aware of others' behaviors within 2 hops [22] and be influenced by indirect ties up to 3 hops [23], we focus this task for pairs of users at social distance 2 and 3.

To investigate how such indirect ties materialize into actual links between users, we compare the performance of three different metrics of indirect tie strength: 1) Jaccard Index (J) [19], 2) Adamic-Adar (AA) [24], and 3) Social Strength (SS), a recently proposed metric [25,26] that quantifies the strength of indirect ties. We note that Jaccard Index and Adamic-Adar consider only the number of shared friends between users, while Social Strength also takes into account interaction intensity.

Social Strength. For completeness, we briefly describe next the Social Strength metric. For measuring the Social Strength of an indirect social tie between users i and m, we consider relationships at n ($n = 2$ or $n = 3$) social hops, where n is the shortest path between i and m. A weighted interaction graph model that connects users with edges weighted based on the intensity of their direct social interactions is assumed. Assuming that $\mathcal{P}_{i,m}^{n}$ 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:

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

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

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

Equation 2 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, and number of co-authored papers. 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 strength 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 (e.g., node i's neighborhood N_i), even in undirected social graphs, the relative weight of the mutual tie will be different from the perspective of each user.

Prediction Task. The link prediction task decides whether the edge (u, v) will form during the observation time. We studied this task on the TF2 and IE datasets. The TF2 network has a timestamp of when a declared relationship was created in Steam Community. However, since for the IE network we do not have formally declared relationships, we use the timestamp of the first recorded face-to-face interaction between two individuals as a proxy for relationship creation.

In TF2, there are $5,984$ 2-hop ($2,475$ 3-hop) pairs that had a relationship formed within the observation time (OT) and $161,561$ 2-hop ($676,863$ 3-hop) pairs who didn't. In IE, there are $1,886$ 2-hop (484 3-hop) pairs that had a relationship formed within OT, and $4,111$ 2-hop ($24,631$ 3-hop) pairs who didn't. This means our datasets are imbalanced with respect to pairs who closed the 2-hop or 3-hop distance or not. There are two common approaches for dealing with unbalanced data classifications: under-sampling [27] and over-sampling [28]. We chose to under-sample pairs of users with no relationships materializing within OT, thus in our experiment they appear at the same empirical frequency as the pairs who formed relationships within OT.

In this prediction task, we used two classic machine learning classifiers: *Random Forest* (RF) and *Decision Tree* (J48). They are tested using tie strength values calculated from the three metrics (Jaccard Index, Adamic-Adar and Social Strength) as features. We used standard prediction evaluation metrics: Precision, Recall, F-Measure and Area Under Curve (AUC) to evaluate the performance of prediction of each classifier and tie strength metric.

4.2 Experimental Results

Table 2 shows the link prediction results of nodes 2 and 3 hops away. Clearly, all three indirect tie metrics demonstrate their power in predicting the formation of links between pairs of non-connected 2-hop users. We note that the AUC reaches 0.77 for the TF2 network using social strength as the metric and J48 as the classifier, and reaches 0.88 for the IE network when using social strength as the metric with random forests as the classifier, greatly outperforming the other two tie strength predictor metrics.

Given that the Jaccard Index and Adamic-Adar metrics are restricted to predictions within 2 hops, we test only the social strength metric for the 3-hop distant link predictions. The results in Table 2 show that while the social strength's effectiveness to predict link formation is reduced, it still manages to properly discriminate between links formed or not by up to about 70% of the time in TF2 and 68% of the time in IE. Overall, while it is expected to see a decrease in performance when we cross the horizon of observability of 2 hops [22], our results show that indirect ties are able to predict the formation of links.

Table 2: Results of link prediction between pairs of n-hop distant users. Only SS is applicable to $n = 3$.

Network	n	Classifier	Metric	Precision	Recall	F-Measure	AUC
TF2	2	RF	SS	**0.71±0.005**	**0.71±0.005**	**0.71±0.006**	**0.76±0.006**
			AA	0.68±0.003	0.67±0.003	0.67±0.003	0.70±0.005
			J	0.67±0.004	0.66±0.003	0.66±0.003	0.70±0.003
		J48	SS	**0.75±0.012**	**0.74±0.008**	**0.74±0.006**	**0.77±0.009**
			AA	0.71±0.004	0.71±0.004	0.71±0.004	0.71±0.006
			J	0.51±0.007	0.51±0.006	0.50±0.008	0.51±0.008
IE	2	RF	SS	**0.81±0.005**	**0.81±0.002**	**0.81±0.003**	**0.88±0.005**
			AA	0.67±0.004	0.66±0.0114	0.66±0.011	0.71±0.002
			J	0.67±0.001	0.66±0.0172	0.66±0.005	0.72±0.002
		J48	SS	**0.84±0.013**	**0.84±0.002**	**0.84±0.002**	**0.87±0.001**
			AA	0.69±0.002	0.69±0.002	0.68±0.003	0.70±0.003
			J	0.69±0.007	0.68±0.005	0.68±0.001	0.68±0.004
TF2	3	RF	SS	**0.653±0.01**	**0.651±0.01**	**0.651±0.01**	**0.709±0.02**
		J48	SS	0.630±0.02	0.627±0.01	0.624±0.01	0.644±0.03
IE	3	RF	SS	**0.659±0.01**	**0.650±0.004**	**0.646±0.004**	**0.682±0.01**
		J48	SS	0.636±0.01	0.633±0.01	0.631±0.01	0.664±0.01

5 Timing of Link Formation

Network dynamics can also be examined from the perspective of *link delays* [29]. If we consider that a link between two nodes is *possible* when all the enabling conditions are met, then the link delay is the time lag between the conditions being met and the link forming. In this section, we investigate if there is a

connection between the strength of a tie of indirectly connected users and the delay the link experiences before it is formed.

5.1 Methodology

Let us consider the toy networks in Figure 1. We define the link formation delay for 2-hop indirect ties (Figure 1a) as:

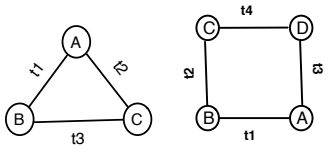
$$\Delta_{(b,c)} = t_{(b,c)} - max\{t_{(a,b)}, t_{(a,c)}\},$$

where $t_{(a,b)}$ is the time when the direct link between two nodes is established. This formulation can also be thought of as the triadic closure delay [29]. Δ thus is a proxy of the "speed" at which two indirectly connected nodes become directly connected: small Δ indicates that the triangle closes quickly, and vice versa.

Similarly, the link formation delay for 3-hop indirect ties (Figure 1b) is:

$$\Delta_{(c,d)} = t_{(c,d)} - max\{t_{(a,b)}, t_{(a,d)}, t_{(b,c)}\}.$$

Although no direct analogue for the 3-hop link formation delay was explored in [29], an n-hop link delay can be considered a form of the general link delay scenario with the restriction that an n-hop path must exist between the two nodes under consideration.



(a) (b)

Fig. 1: (a) B and C have a 2-hop relationship before t_3, since $t_1, t_2 < t_3$, and a 1-hop relationship thereafter. (b) C and D have a 3-hop relationship before t_4, since $t_1, t_2, t_3 < t_4$, and a 1-hop relationship thereafter.

To measure the strength of indirect ties, we employ the social strength metric to quantify the strength of a social connection between indirectly connected nodes. We are primarily interested in whether the latent tie strength between indirectly connected nodes corresponds to different delays in a direct connection forming. Intuitively, if the strength of a user's indirect tie is stronger than any of the user's strong direct ties, we consider it a strong indirect tie. Because we have no information regarding the strength of a direct tie (other than the edge weight), we consider an indirect tie of a's as strong if its strength is larger than the minimum/average/maximum weight of all of a's direct edges. These alternative criteria are formally presented below. (We note that the social strength metric is asymmetric, i.e., $SS(a,b) \neq SS(b,a)$):

$$\text{[C-min]: } SS(a,b) \geq \min_{i \in Neigh(a)} [NW(a,i)] \text{ or } SS(b,a) \geq \min_{a \in Neigh(i)} [NW(i,a)]$$

$$\text{[C-mean]: } SS(a,b) \geq \frac{\sum_{i \in Neigh(a)} [NW(a,i)]}{size(Neigh(a))} \text{ or } SS(b,a) \geq \frac{\sum_{a \in Neigh(i)} [NW(i,a)]}{size(Neigh(i))}$$

$$\text{[C-max]: } SS(a,b) \geq \max_{i \in Neigh(a)} [NW(a,i)] \text{ or } SS(b,a) \geq \max_{a \in Neigh(i)} [NW(i,a)]$$

In each criterion, $NW(a,b)$ is the normalized weight of the edge between nodes a and b, and the normalization is conducted by the total weight of node a's edges. If an indirect tie (a,b) satisfies the conditions for a given criterion, it is marked as a strong indirect tie; otherwise it is marked as a weak indirect tie. Table 3 summarizes the tie classification results when these criteria are applied to the networks TF2 and IE.

Table 3: The statistics of 2- and 3-hop indirect ties in TF2 and IE networks where ties are divided into strong and weak ties under three criteria.

Dist.	Network	Tie classification criterion	# strong ties	# weak ties
2	TF2	C-min	6,868	164
2	TF2	C-mean	5,470	1,562
2	TF2	C-max	2,780	4,252
3	TF2	C-min	2,351	90
3	TF2	C-mean	297	2,144
3	TF2	C-max	12	2,429
2	IE	C-min	1,555	42
2	IE	C-mean	1,235	344
2	IE	C-max	715	882
3	IE	C-min	193	258
3	IE	C-mean	11	440
3	IE	C-max	0	451

5.2 Experimental Results

We use the TF2 and IE networks described in Section 3 to analyze link delays when examining 2- and 3-hop indirect ties. We compare the link delay of weak and strong ties classified by the previously defined criteria. For TF2, we use *days* as the time unit, but for IE we use *minutes* due to the ephemeral nature of its face-to-face interactions.

The link delay distributions are plotted in Figure 2, where we see that pairs with strong indirect ties formed direct links with shorter delay than those with weak indirect ties. We note that strong ties formed their link with less delay than weak ties throughout all scenarios and when the tie is stronger, its link formed even quicker. For example, when using 3-hop indirect ties in TF2 and criterion *C-max* for classifying strong vs. weak, 33% of strong indirect ties formed a direct link within a day, compared to only 7% for weak indirect ties. In contrast, over 40% of weak indirect ties formed direct links with a large delay (over 60 days). Overall, these results indicate several things. First, when indirect ties are stronger, there is an increased chance for them to establish a link quicker. Second, even quantifying the strength of the tie from 3-hops away, strong indirect ties led to faster link creation.

6 Interaction Intensity along Newly Formed Links

A key characteristic of social interactions is their continuous change, and this change is likely to affect user behavior related to network dynamics. E.g., fre-

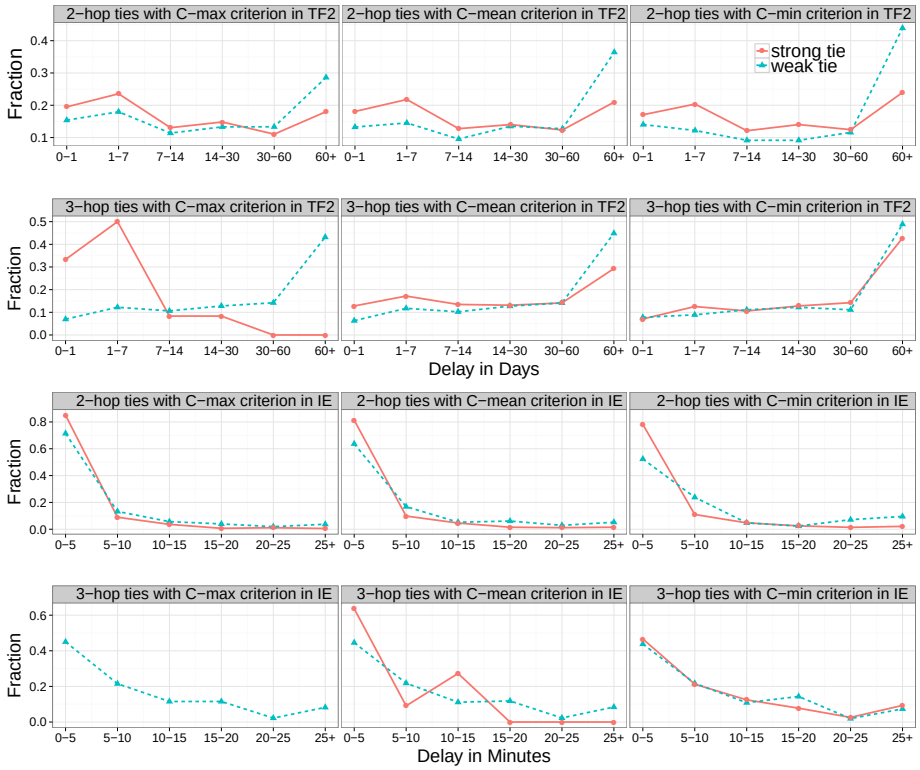


Fig. 2: Link delay comparison between 2- and 3-hop strong vs. weak ties in TF2 and IE. Note that for the IE network, when 3-hop ties are divided by criterion *C-max*, no strong ties exist.

quent interactions lead to the formation of new links, and by interacting with each other, information can be disseminated in the network. Thus, we believe the changes in the interactions between nodes previously connected by indirect ties also can predict the dynamic status of the network.

Note that among all four datasets introduced in Section 3, only the online game social network (TF2) supplies a timestamp for each friendship formation and interaction. More importantly, because gamers can play with each others without being declared friends in Steam Community OSN, we can measure interaction intensity in the absence of a declared relationship. Thus, in the following our analysis is based on TF2 network.

We analyze the intensity of user interactions before and after a pair of users, who are 2- and 3-hop away, form a new edge. Figure 3 shows that in both scenarios (2 and 3 hops), more pairs of users have interactions after their link formation than before the link formation. For example, 54% pair of users have no interactions before they establish an edge with each other, while this number decreases to 17% after a 2-hop indirect tie is closed with a direct tie. This result shows

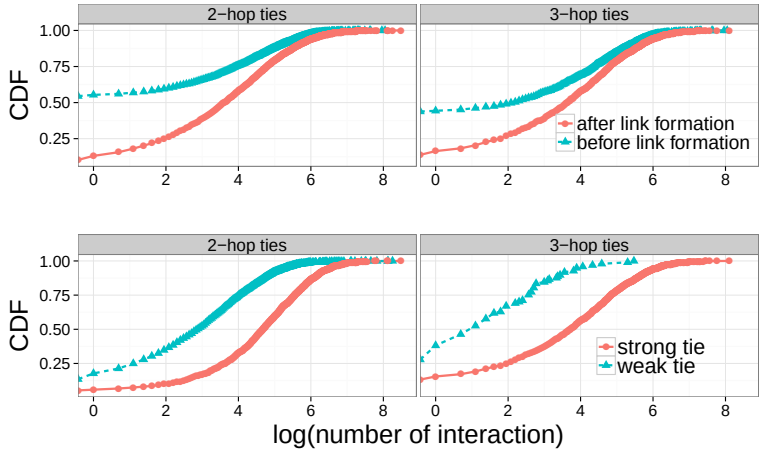


Fig. 3: Interaction intensity before vs. after 2- and 3-hop link formation in TF2, and strong vs. weak ties' (identified by the C-min criterion) interaction intensity after link formation of TF2

that after users form a direct link, their interactions not only continue but also increase, implying that users actively maintain their newly formed relationship.

As a further step, we investigate the difference of interaction intensity between nodes previously connected via strong vs. weak indirect ties after forming their direct links. We use the *C-min* criterion introduced in Section 5 to classify indirect ties into strong and weak. Figure 3 plots interaction intensity after link formation. The figure shows that strong indirect ties lead to direct ties with more interactions than weak indirect ties do.

7 Predicting Information Diffusion Paths

Information diffusion is a fundamental process in social networks and has been extensively studied in the past (e.g., [30,31,32,33,34]). In fact, some studies have shown that the evolution of a network is affected by the diffusion of information in the network [33] and vice versa [32]. Our results from the previous sections show that indirect ties affect the process of network evolution. 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 step t, can we predict which other users will receive this information at time step $t+n$ ($n \geq 2$)? I.e., if we know someone who received the information at t_0, then can we directly predict the infected users at t_n ($n \geq 2$) instead of step-wise (e.g., at t_1)?

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 offi-

cial 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 several indirect-tie metrics to predict information diffusion paths.

7.1 Experimental Setup

The strength of an indirect tie decreases with the length of the shortest path between the two individuals. This has been quantitatively observed by Friedkin [22], who concluded that people's awareness of others' performance decreases beyond 2 hops. Three degrees of influence theory, proposed by Christakis et al. [23], 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 a variety of social networks examined [35,36] and in accordance with these observations, 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 with the knowledge from t_0.

Diffusion Simulation. As ground truth, we applied the basic and widely studied *Linear Threshold (LT)* diffusion model [37] to simulate a diffusion process, i.e., which nodes are affected during each time step.

The LT model is a threshold-based diffusion model where nodes can be in one of two states: active or inactive. We say a node has accepted the information if it is active, and once active, it can never return to the inactive state. In the LT model, a node v is influenced by each of its neighbors $Neigh_v$ according to an edge weight $b_{v,w}$. Each node v chooses a *threshold* θ_v that is randomly generated from the interval [0,1]. The diffusion process is simulated as follows: first, an initial set of active nodes A_0 is chosen at random and these are the seed nodes. Then, at each step t, all nodes that were active in step $t-1$ remain active, and we activate any node v for which the total weight of its active neighbors is at least θ_v, that is $\sum_{w \in Neigh_v} b_{v,w} \geq \theta_v$. Thus, the threshold θ_v intuitively represents the different latent tendencies of nodes to adopt the behavior exhibited by neighbors, and a node's tendency to become active increases as more of its neighbors become active. The input to the simulation is a weighted graph where edge weights represent the intensity of interactions between nodes (n.b., the LT model considers only the status of a node's directly connected neighbors).

We controlled the effectiveness of the diffusion by gradually changing the upper bound of the thresholds applied on the nodes to simulate different diffusion processes; from almost no diffusion to fully dissemination to all nodes in the graph. To do so, we set a threshold $\theta_v = random(0,1)/w$ where w is empirically selected based on the range of edge weights in each of the tested networks, i.e., w in the range of [1-10] for the CA-I, [1-30] for the CA-II and [1-60] for the TF2.

Predicting Diffusion Paths via Indirect Ties. Once we generate the ground truth from the LM model, we then use the strength of indirect ties to predict the

path of diffusion. To measure the strength of indirect ties, we also employ social strength, Adamic-Adar and Jaccard metrics introduced in Section 4.1 where the social strength metric considers the edge weight while Adamic-Adar and Jaccard only consider the neighborhood overlap. We calculate the strength of indirect tie values between the seed and its n-hop nodes, then convert the values to a social rank. Each user has a rank list for all her n-hop nodes according to the strength of the indirect tie value between the user and the node.

After obtaining social ranks, we need a cut-off threshold to decide whether or not a node's n-hop nodes will be active at t_{0+n}(n=2 or 3). Our strategy requires that the social ranks from information recipient's perspective must be high, e.g., $socialrank_n(A, B)$ ranks among the top 10% of user A's contacts. Then, the cut-off threshold can classify a node's n-hop nodes into two categories: active or inactive at t_{0+n}(n=2 or 3). The intuition of this cut-off is that users will likely believe the information from their "closest" social ties. The cut-off threshold can be calculated as $\theta_{pred} = |Neigh_{nhops}|/q$ where q is empirically selected to have an inversely proportional relationship to w, which decides the diffusion process from almost no diffusion to full dissemination to all nodes. In other words, when no diffusion happens the θ_{pred} should be small enough to select the strongest indirect ties while in a fully diffused scenario a larger θ_{pred} is needed to cover a large portion of indirect ties.

7.2 Results and Evaluation

In literature, co-authorship networks capture many general features of social networks [38] and have been studied in information cascades [37], and diffusion dynamics have been observed in online game social networks [39,40]. Therefore, in our experiments, we use the three datasets—CA-I, CA-II and TF2—as described in the previous section. To better demonstrate indirect ties' effective power on inferring diffusion processes, we compare indirect tie metrics with a baseline method, which randomly selects a information recipient's 2 and 3-hop friends to accept the information.

We compare the prediction results with the ground truth obtained from the diffusion simulation to verify the effectiveness of indirect ties in predicting diffusion paths. We evaluate our method using accuracy, sensitivity and specificity [41]. Figures 4 and 5 depict the prediction results in a 2- and 3-hop social distance, respectively. We see that for both 2- and 3-hop path predictions, overall the accuracies of indirect tie metrics are higher than the baseline's, reaching a maximum of 0.90 with social strength metric in 2-hop path predictions. Also, the accuracies of the three indirect tie metrics in all cases are always higher than 0.56, and social strength outperforms the other two metrics in most of the scenarios. Although 3-hop predictions (generated by the Social Strength metric) show decreased sensitivity, specificity and accuracy compared to 2-hop results, they remain above 0.64. It is important to note that these three networks have very different network structure (from sparse to dense), yet the performance of indirect tie metrics are consistently higher than the baseline in all three networks and for different diffusion thresholds. From these results, we conclude that indirect ties can be used in the prediction of information diffusion, i.e., along which

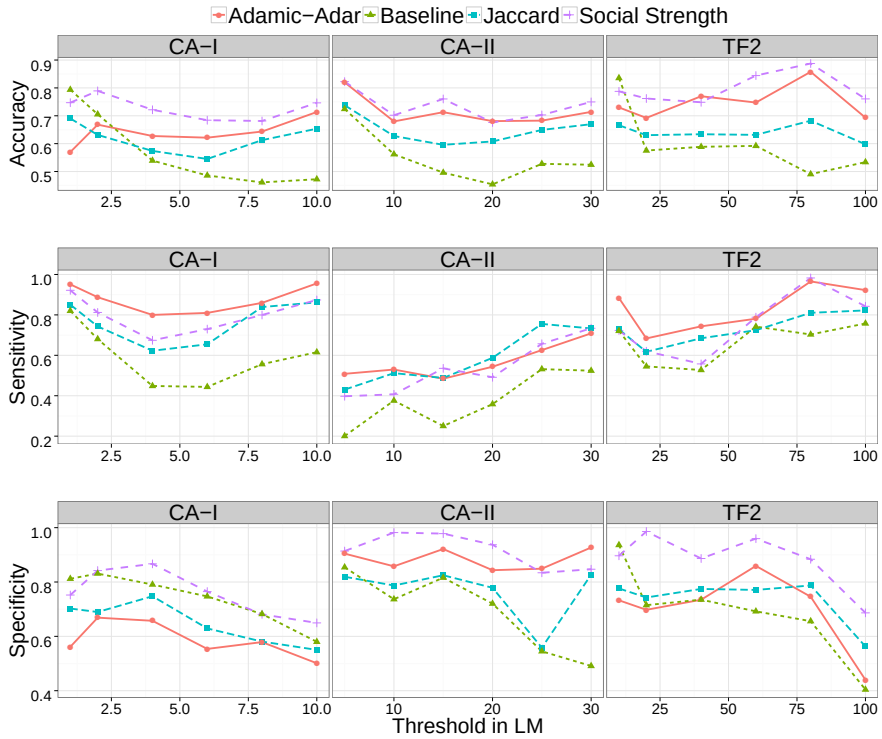


Fig. 4: Performance of different measures of strength for 2-hop indirect ties, in the prediction of information diffusion paths in the networks CA-I, CA-II and TF2.

paths information will propagate and which users will be activated, at least 2-3 steps before a susceptible node is even in contact with an infected node.

8 Summary and Discussions

In this paper, we empirically examine the predictive power of indirect ties in network dynamics. By using four real-world social network datasets and three indirect measurements, we empirically show that indirect ties can be used for predicting the newly formed edges and the stronger an indirect tie is, the quicker the tie will form a link. In addition, strong indirect ties correlate to more interactions, and the interaction has the tendency to be continued after the link formed. Finally, we show that indirect ties can also be used for predicting information diffusion paths in social networks.

This is our first step to investigate the influences of indirect ties on network dynamics. In the future, we will further study the effects of indirect ties on information diffusion paths with various diffusion models and real-world cascades.

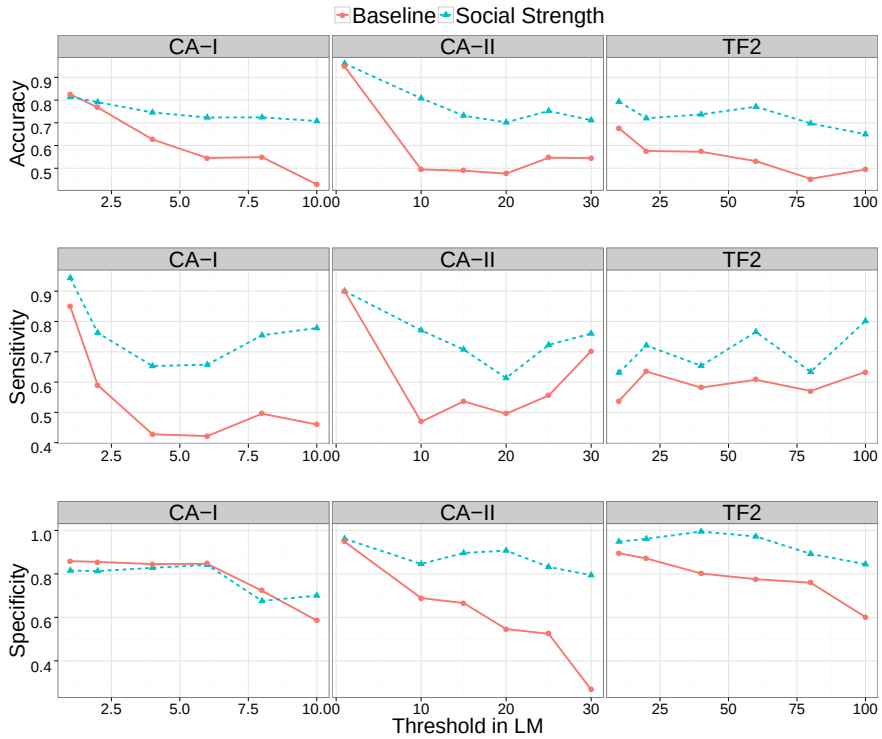


Fig. 5: Performance of different measures of strength for 3-hop indirect ties, in the prediction of information diffusion paths in the networks CA-I, CA-II and TF2.

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. Z. Li and H. Shen, "Soap: A social network aided personalized and effective spam filter to clean your e-mail box," in *INFOCOM, Proceedings IEEE*, 2011.
2. C. Basu, H. Hirsh, and W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation," in *AAAI/IAAI*, 1998, pp. 714–720.
3. J. Li and F. Dabek, "F2F: reliable storage in open networks," in *Proceedings of the 4th International Workshop on Peer-to-Peer Systems (IPTPS)*, 2006.
4. I. Kahanda and J. Neville, "Using transactional information to predict link strength in online social networks." in *ICWSM*, 2009.

5. E. Gilbert and 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.
6. M. S. Granovetter, "A study of contacts and careers," *Cambridge, Mass. Harvard*, 1974.
7. M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.
8. M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, 1973.
9. A. Rapoport, "Spread of information through a population with socio-structural bias: I. assumption of transitivity," *The bulletin of mathematical biophysics*, vol. 15, no. 4, pp. 523–533, 1953.
10. G. Kossinets and D. J. Watts, "Origins of homophily in an evolving social network1," *American Journal of Sociology*, vol. 115, no. 2, pp. 405–450, 2009.
11. A. M. Kleinbaum, "Organizational misfits and the origins of brokerage in intrafirm networks," *Administrative Science Quarterly*, vol. 57, no. 3, pp. 407–452, 2012.
12. L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: memberships, growth, and evolution," in *In the International conference on Knowledge Discovery and Data mining (KDD)*, 2006.
13. A. Patil, J. Liu, and J. Gao, "Predicting group stability in online social networks," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1021–1030.
14. J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," in *ICWSM*, vol. 10, 2010, pp. 355–358.
15. J. Blackburn and A. Iamnitchi, "Relationships under the microscope with interaction-backed social networks," in *1st International Conference on Internet Science*, 2013.
16. J. Blackburn, N. Kourtellis, J. Skvoretz, M. Ripeanu, and A. Iamnitchi, "Cheating in online games: A social network perspective," *ACM Transactions on Internet Technology (TOIT)*, vol. 13, no. 3, p. 9, 2014.
17. L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. V. den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 166–180, 2011.
18. J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.
19. D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.
20. L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.
21. R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *19th International Conference on World Wide Web*, Raleigh, NC, USA, 2010, pp. 981–990.
22. N. E. Friedkin, "Horizons of observability and limits of informal control in organizations," *Social Forces*, vol. 62, no. 6, pp. 54–77, March 1983.
23. N. A. Christakis and J. H. Fowler, *Connected: The surprising power of our social networks and how they shape our lives*. Hachette Digital, Inc., 2009.
24. L. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211—230, 2003.

25. N. Kourtellis, "On the design of socially-aware distributed systems," Ph.D. dissertation, University of South Florida, 2012.
26. X. Zuo, J. Blackburn, N. Kourtellis, J. Skvoretz, and A. Iamnitchi, "The power of indirect ties in friend-to-friend storage systems," in *14th IEEE International Conference on Peer-to-Peer Computing*, September 2014.
27. M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *ICML*, vol. 97, 1997, pp. 179–186.
28. N. V. Chawla, K. W. Bowyer, H. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
29. M. Zignani, S. Gaito, G. P. Rossi, X. Zhao, H. Zheng, and B. Y. Zhao, "Link and triadic closure delay: Temporal metrics for social network dynamics," in *ICWSM*, 2014.
30. M. Yildiz, A. Scaglione, and A. Ozdaglar, "Asymmetric information diffusion via gossiping on static and dynamic networks," in *Decision and Control (CDC), 49th IEEE Conference on*, Dec 2010, pp. 7467–7472.
31. A. Guille and 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.
32. E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW, 2012, pp. 519–528.
33. L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and 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*, ser. KDD, 2013, pp. 356–364.
34. A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, July 2013.
35. J. H. Fowler, N. A. Christakis, and D. Roux, "Dynamic spread of happiness in a large social network: longitudinal analysis of the framingham heart study social network," *BMJ: British medical journal*, pp. 23–27, 2009.
36. N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England journal of medicine*, vol. 357, no. 4, pp. 370–379, 2007.
37. D. Kempe, J. Kleinberg, and Éva Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM International Conference on Knowledge Discovery and Data Mining*, ser. KDD. ACM, 2003, pp. 137–146.
38. M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.
39. X. Wei, J. Yang, L. A. Adamic, R. M. de Araújo, and M. Rekhi, "Diffusion dynamics of games on online social networks," in *Proceedings of the 3rd conference on Online Social Networks*. USENIX Association, 2010, pp. 2–2.
40. J. Blackburn, R. Simha, N. Kourtellis, X. Zuo, M. Ripeanu, J. Skvoretz, and A. Iamnitchi, "Branded with a scarlet "c": cheaters in a gaming social network," in *Proceedings of the 21st International conference on World Wide Web*, 2012.
41. T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.