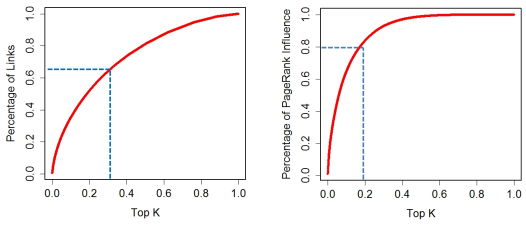
Predicting Node Degree Centrality with the Node Prominence Profile
Supporting Information

Yang Yang[1], Yuxiao Dong[1], and Nitesh V. Chawla[11]

[1]Department of Computer Science and Engineering, University of Notre Dame

October 9, 2014

[1]nchawla@nd.edu



(a) Degree Centrality Ranking (b) PageRank Centrality Ranking

Figure 1: Pareto Principle.

1 Data and Problems

1.1 Datasets

In this paper we examine our approaches and perform our analysis on four social networks. The **Condmat** network [25] is extracted from a stream of 19,464 multi-agent events representing condensed matter physics collaborations from 1995 to 2000. Based on the **DBLP** dataset from [27] we attach timestamps for each collaboration and choose 3,215 authors who published at least 5 papers. **Enron** dataset [28] contains information of email communication among 16,922 employees in Enron Corporate from 2001.1.1 to 2002.3.31. The **Facebook** dataset is used by Viswanath et al. [29], which contains wall-to-wall post relationship among 11,470 users between 2004.10 and 2009.1.k

1.2 Problem Definition

Network evolution is usually reflected in changes of node's centrality. First, the social network evolution impacts the centrality (relative importance in the network) of node; in addition, the node also affects its local neighbourhood and beyond (via link formation or link dissolution). In order to give insights into the network dynamics, we provide several definitions and formulate the concrete problems for the ease of evaluation and comparison.

On a global level, important nodes (high centrality) have intrinsically higher strength of impact than others due to the network topology. Through our study, we have found that a small number of nodes occupy large portion of network resources. For example, in Figure 1(b) top 20% (ranked by PageRank) nodes occupy about 80% *PageRank* centrality in DBLP network. This satisfies *Pareto Principle* (also known as 80-20 rule) [9]. To better understand and model the effects of network evolution on node's centrality, we partition nodes into two sets *important nodes* and *non-important nodes*. Based on *Pareto Principle*, their definitions are given as follows:

Definition 1. Important Node In a network $G = (V, E)$ a node v is a *important node* under centrality measurement $\mathbb{M}$ if and only if $\frac{|\{u|\mathbb{M}(u)\leq\mathbb{M}(v)\}|}{|V|} \geq 0.8$.

Definition 2. Non-Important Node In a network $G = (V, E)$, a node v is a *non-important node* under centrality measurement $\mathbb{M}$ if and only if $\frac{|\{u|\mathbb{M}(u)>\mathbb{M}(v)\}|}{|V|} \geq 0.2$.

In following sections we denote the set of *important nodes* as **IN** and the set of *non-important nodes* as **NIN**.

Prediction of centrality can lead us to infer influence, importance and/or success of a given individual in a social network. We use the popular degree centrality as a metric in this paper (various studies have found centrality measures to be correlated). As we postulated, an arguably generic approach for the network evolution analysis should have the predictability of a node's degree centrality in the future. Therefore we formulate a concrete task, the degree centrality prediction problem,

where we can directly evaluate different approaches and facilitate our findings of the underlying principles. The associated definitions are as follows:

Definition 3. Time-varying Network The time-varying network at time t is denoted as $G_t = (V, E, T_V, T_E)$, where V is the set of nodes and E is the set of links among nodes, T_V is the set of arriving time of all nodes and T_E is the time log of all links.

Definition 4. Degree Centrality Prediction In a time-varying network $G_t = (V, E, T_V, T_E)$ the degree centrality prediction task is, for the set of nodes $V_t = \{v | v \in V, T_V(v) = t, v \notin \mathbf{IN}_t\}$, where $\mathbf{IN}_t$ is the set of *important nodes* measured in network G_t. How reliably can we infer whether a node v ($v \in V_t$) will belong to the set of $\mathbf{IN}_{t+\Delta T}$?

In order to demonstrate the discrimination of different principles, we consider the newly joining nodes as the prediction instances and ΔT is selected large enough for the node degree centrality evolution. For example in DBLP we consider nodes that join the collaboration network in recent 2 years and predict whether they will become important after 5 years. In this paper we concentrate on predicting nodes' future degree centrality. We use the popular degree centrality as a metric in this paper, due to the fact that various studies have found centrality measures to be correlated.

These concrete problems provide us quantitative and microscopic views of network evolution, which also make it convenient for principles comparison. Additionally, we also study whether the learned predictors can generalize across different domains of social networks for the problem defined above. This provides us rigorous and empirical views of the network evolution problem.

2 Node Prominence Profile

An important fraction of network dynamics locates in the process of degree centrality evolution. A generic and effective measurement should be able to infer the degree centrality evolution trend, and aid in predicting the potential degree centrality of a node in the future. We first introduce the current state-of-art of centrality measures and discuss their limitations. In addition, we validate the fundamental principles associated, and introduce our framework - **node prominence profile** which optimizes trade-offs between preferential attachment and triadic closure. Finally, based on experiments we unveil the interactions between node's degree centrality and its position, which are well reflected in our framework.

2.1 Current State of The Art

Many centrality measures are proposed for identifying important (high centrality) nodes in a network, such as *degree centrality*, Pagerank [5], Betweenness [6], and Closeness [7].

2.1.1 Limitations of Current Methods

Although these methods are proved to be effective in centrality quantification and measurement, they are inherently lack of predictability. First, most of these measures assign a value to each node, which leads to the loss of information; second, much research has focused on describing the centrality and ignore the position of nodes in the network. To summarize, even though these existing centrality measures are good at evaluating consequences of network evolution, they have limitations in describing the node's future degree centrality.

2.1.2 A Case of Local Sub-structure

We posit importance of a node relies on its neighborhood. Thus the future importance of a node may be a function of the sub-structure surrounding the node at time t. We have several canonical

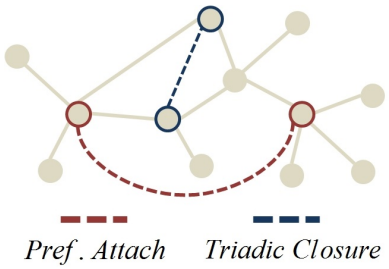


Figure 2: Preferential Attachment vs. Triadic Closure. Based on the principle of preferential attachment two red nodes are most likely to be connected in future; while the triadic closure principle suggests that link between two blue nodes.

examples to support this proposition. First, based on the *PageRank* heuristic: centrality of a node is indicated by the number of connections or links to that node; second, Burt [20] proposed the concept of *structural hole*: a node's success often depends on their access to local bridges. Both of these examples imply that the position of node within a social network is important. This leads us to investigate the value of a node's position within the local sub-structures and the impact on its future degree centrality, which inspires the development of our framework.

2.2 Preferential Attachment and Triadic Closure

Despite the well known macroscopic scaling in social networks, such as *power-law* degree centrality distribution [19], *attachment kernel* [36] and *clustering coefficient* as function of node degree centrality [19], it is undecided whether there is a common mechanism underlying these macroscopic laws [32] [33]. With the evidence that the *preferential attachment* process [19] is just one dimension of network evolution, much recent research has extended the *preferential attachment* principle by local sub-structure evolution rules [38] [39]. Li et al. [39] and Jin et al. [38] proposed that an individual's link formation significantly relies on its neighbors. In the work of [31], Granovetter proposed that a "forbidden" triad (left in Figure 3) is most unlikely to occur in social networks, which means that the probability of a new link to close "forbidden" triad is higher than the probability of link between two randomly selected nodes. The principle of *triadic closure* is demonstrated to be relevant for social network evolution in many works [32] [38] [39] [44]. Obviously these two principles propose two distinct mechanisms of network evolution and none of them can act as a single origin of network evolution. In preferential attachment new links are made preferentially to high degree centrality nodes while in triadic closure new links are generated to close "forbidden" triad (Figure 2). We are interested to know whether there is an effective combination of these two principles.

The principles of *preferential attachment* and *triadic closure* have been empirically demonstrated to be relevant (not as a single origin) for macroscopic scaling laws in the work of [37] [41] [42] [43], expressly or implicitly. As the fact that these principles are underlying the social network macroscopic scaling laws, we are interested to know whether these principles are valid to answer the microscopic problems in social network dynamics, such as *importance prediction*. Our work is different from the work of [37] and [40], Leskovec et al. [37] employ *triadic closure* to reproduce the observed macroscopic laws of social networks and Lou et al. [40] investigated how a reciprocal link is developed and how relationships develop into triadic closure.

2.2.1 Triadic Closure Effect on Degree Centrality Evolution

The effect of *preferential attachment* on the degree centrality evolution is obvious and evident. Here we explore the effect of *triadic closure* on the degree centrality evolution. The quantity of *triadic*

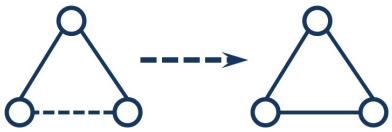
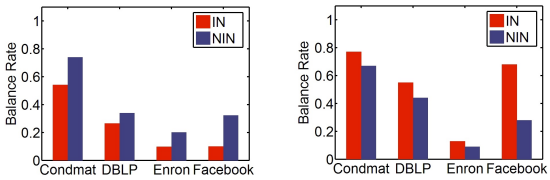


Figure 3: Triadic Closure



(a) Before Evolution (b) After Evolution

Figure 4: Structural Balance Rate

closure (or structural balance) is usually defined as below [38]:

$$\text{balance rate} = \frac{3 \times \text{number of closed triads}}{\text{number of connected triads}} \tag{1}$$

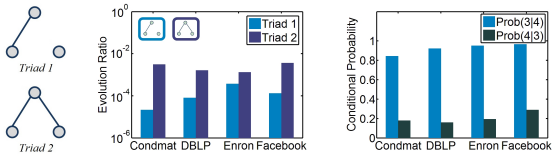
where connected triad is the left triad in Figure 3 and closed triad is the right triad in Figure 3 respectively. By studying the sub-networks among important (high degree centrality) or non-important nodes, we observe that initially the future important nodes sub-network has a lower balance rate than the future non-important sub network, while after long enough evolution the former sub-network forms a more balanced structure (Figure 4). There are several implications:

- There exists connections between the *triadic closure* and the degree centrality evolution. In addition, as discussed above, new links are more likely to form between nodes located in an imbalanced sub-network;
- The initial sub-network where future important nodes are located is more imbalanced than that of future non-important nodes, position of node can be indicative of its future degree centrality.

To some extent this implies the effect of *triadic closure* on the degree centrality evolution mechanism.

Evolution Ratio In the process of network evolution, one type of triads is possible to evolve into another type (Figure 3). For a specific type of triads (i.e., triad 1), we calculate what percentage of them evolve and denote that as the *evolution ratio*. A toy example is given in Figure 6. In this figure we can see after the evolution of network, $\frac{2}{3}$ triad 2 sub-structures evolve, and the evolution ratio of triad 2 in this process is $\frac{2}{3}$.

As suggested in the principle of *triadic closure*, a "forbidden" triad is more likely to attach new links. In order to demonstrate that position is important for evolution, we provide the evolution ratio of two types of triads in Figure 5 (a). We can see that the "forbidden" triad (*triad 2*) has much higher probability to form a new link than the disconnected sub-structure *triad 1*. This implies, nodes in different triads have different probabilities to develop importance. This leads us to an important conclusion: the positions of nodes in sub-structures determine their future orbits in both essential evolution elements for the importance evolution. This observation leads us to develop our framework called, the Node Prominence Profile.



(a) (b) (c)

Figure 5: Structural Balance Statistics

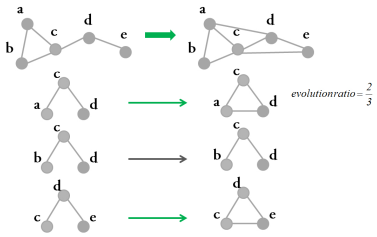


Figure 6: Triad Evolution Ratio Toy Example

Significance Validation In Figure 4 we show that initially the future important nodes (**IN**) sub-network has a lower balance rate than the future non-important (**NIN**) sub-network, while after long enough evolution the former sub-network forms a more balanced structure. In order to validate the significance of these findings, we design the following experiments. In the network we have two categories of nodes, **IN** and **NIN** (important nodes and non important nodes). Our purpose is to demonstrate that (i) **IN** sub-network has a lower balance rate than **NIN** sub-network before evolution and that (ii) IN sub-network has a higher balance rate than the **NIN** sub-network after evolution. A straightforward way to measure and validate this is to compare the balance rate differences between IN sub-network and **NIN** sub-network to the null model where the type of nodes is randomized (the numbers of **IN** and **NIN** do not change, but the type (**IN** or **NIN**) is randomly assigned to nodes). To measure statistical significance of balance rate differences between **IN** and **NIN**, we compare each real network (Condmat, DBLP, Enron, and Facebook) to 10,000 surrogate networks where the type (**IN** or **NIN**) of the nodes was randomly shuffled, leaving the topology of the network intact.

For example, in Condmat network the balance rate difference between **IN** and **NIN** is $D_{real} = 0.230$ before evolution. In each round of simulation, we first randomly assign the type of nodes and then we calculate the balance rates of new **IN** sub-network and **NIN** sub-network. Trivially we have a simulated balance rate difference D_{sim}. After 10,000 round of simulations, statistical significance of balance rate difference is measured by the Z score:

$$Z = \frac{D_{real} - \overline{D}_{sim}}{std(D_{sim})}$$

The Z scores for four real-world networks before and after evolution are provided in Table 1 and Table 2 correspondingly. In Table 1 and Table 2 we present Z-scores with four different thresholds (i.e., 10%, 20%, 30%, and 50%). As we have observed in Figure 4, the balance rate of **IN** sub-network is smaller than that of **NIN** sub-network. Thus, the balance rate difference between **IN** and **NIN** $D_{real} < 0$. Clearly, most balance rate differences between **IN** and **NIN** are significantly under-represented (marked by blue color). This means our observation that balance rate of **IN** sub-network is smaller than that of **NIN** sub-network is statistically significant, even if we set the threshold to 50%. The only exception is, in Enron network where we set the threshold as 50%, the balance rate of **IN** sub-network is not substantially smaller ($Z = -0.298$). Additionally the

significance levels of other thresholds (10%, 20%, and 30%) in Enron are lower than other networks (approximately 2 sigmas). This indicates that the triadic closure effect is not that significant in Enron network, which well explains why *Triadic Closure* based method has much worse performance in Enron (see Table 7, method **TC**).

Similarly we also provide the significance validation for the network after evolution. The real balance rate difference $D_{real} > 0$. If we observe that the real balance rate is significantly larger than the random cases, we can claim that the balance rate of **IN** sub-network is significantly larger than that of **NIN** sub-network. In Table 2 we can observe that in most cases the observation that **IN** sub-network is more balanced than the **NIN** sub-network is significant. In Table 2 when we set the threshold to 50% in Condmat, the observation is less substantial ($Z = 1.659$). Additionally in Enron when we set the threshold to 30% and 50%, the observation is also not significant. This further confirms that the triadic closure effect is not comparably significant in Enron network. Although in some cases the observation is not substantial, our experiments already provide enough evidence that triadic closure has effects on the degree centrality evolution. Initially future important nodes are located in a more imbalanced sub-network than that of future non-important nodes, and the sub-network of important nodes evolves to be more balanced or comparably balanced than non-important nodes.

Table 1: Significance Validation of Social Balance Rates Before Evolution. Red color indicates the significance level is less than 2 sigmas (H_0 rejected at 2 sigmas).

Nodes Partition	Z-scores			
Percentile	Condmat	DBLP	Enron	Facebook
10%	-6.696	-4.913	-2.884	-6.171
20%	-6.023	-2.096	-2.501	-11.253
30%	-7.955	-4.220	-1.972	-10.568
50%	-4.053	-9.859	-0.298	-2.744

Table 2: Significance Validation of Social Balance Rates After Evolution. Red color indicates the significance level is less than 2 sigmas (H_0 rejected at 2 sigmas).

Nodes Partition	Z-scores			
Percentile	Condmat	DBLP	Enron	Facebook
10%	3.697	7.133	3.705	14.626
20%	3.515	5.682	4.400	12.241
30%	4.708	9.875	1.794	3.063
50%	1.659	4.093	-0.578	2.139

2.3 Node Prominence Profile

Motivated by the above analysis, we start our investigations from the principle of *triadic closure*. Based upon the principle of *triadic closure*, an individual will try to close a "forbidden" triad that it has, for example in Figure 3 a "forbidden" triad is likely to evolve as a closed triad. See examples of all possible triads in Figure 7(a) and Figure 7(b). The number labeled on the edge describes whether two nodes have relation, for instance '1' can state that two actors are friends while '0' means they are non-friends.

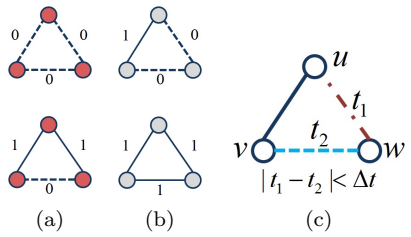


Figure 7: Triads and User Influential Probability.

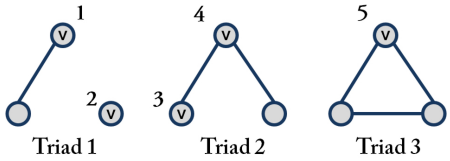


Figure 8: Node Prominence Profile

Such kind of triad evolution has very nice characteristics, firstly it leads to the formation of the link, and additionally it also increases the degree centrality of node. Thus, different positions of a node in corresponding triads can be indicative of degree centrality. This satisfies our proposition that the prominence of a node not only includes its centrality but also its position in local neighborhood. As we have discussed above, the position of node within substructures could provide us insights into the principles underlying the network evolution. To that end, in Figure 8 we enumerate all possible five positions in the triad sub-structures for further study. We are interested to know that the consequences of preferential attachment and triad closure on the network evolution; second, we want to validate our proposition made in the above section and seek a solution which optimizes trade-offs between two distinct principles.

Based on our discussions above, we introduce our framework-*Node Prominence Profile* for the degree centrality evolution analysis. Formally, the *Node Prominence Profile* is defined as follows:

Definition 5. Node Prominence Profile *Node Prominence Profile* for a node v, written as NPP(v), is a vector describing the occurrence frequencies of node v in five different positions in three isomorphic substructures.

In order to analyze the generality and effectiveness of existing centrality measures and our method, we design an experiment to identify their correlation with node latent degree centrality. The evidence that our framework combining two principles well will be provided later.

An Example of Node Prominence Profile In Figure 8, there are three sub-structures and five automorphism positions. To compute the node prominence profile for an individual node v, we need to find out all these sub-structures where node v is located. Then we count how many times node v occurs in each automorphism position.

An example is given in Figure 9. In Figure 9, there is a network with five nodes and five edges; and in total there are 6 sub-structures described in the definition of Node Prominence Profile (Triad 1, Triad 2, and Triad 3 in Figure 8). Among these six sub-structures, node s is located in different positions. For example, node s is in position 1 for twice and in position 2 for zero times. In this way, after finding out all associated sub-structures, we just count how many times node s has shown up in each position. And finally, we can calculate the node prominence profile of node s as $NPP(s) = (2, 0, 1, 2, 1)$.

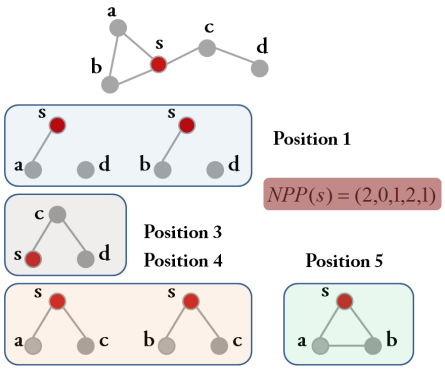


Figure 9: Toy Example of Node Prominence Profile. To compute the node prominence profile for an individual node v, we need to find out all these sub-structures where node v is located. And then we count how many times node v occur in each automorphism position. In this figure, we calculate the node prominence profile of node s in a sample network (five nodes and five edges). Node s never show up in position 2 in triad 1, thus the corresponding value of position 2 in node prominence profile is zero.

Experimental Setup for Significance Validation For a time-varying network at time t G_t, we extract the set of nodes whose arriving time is t and then compute their centrality measures based on the topology of G_t. At time $t + \Delta T$ for the network $G_{t+\Delta T}$ we classify the set of nodes into important (high degree centrality) set $\mathbf{IN}_{t+\Delta T}$ and non-important (low degree centrality) set $\mathbf{NIN}_{t+\Delta T}$ based on the topology of $G_{t+\Delta T}$. In order to demonstrate the discrimination of two principles, ΔT is selected large enough for the node degree centrality evolution. As we know when ΔT is small, the degree centrality prediction problem will be easy. Here we extract new arriving nodes as our prediction candidates, because existing nodes are well evolved and much easier to predict. In this way for each new arriving node we have its centrality measures and triad position counts measured at time t, and also we have its corresponding degree centrality at time $t + \Delta T$. We employ Wald test [45] to evaluate the relationship between each feature (centrality measures or **NPP** position counts measured at time t) and corresponding future degree centrality (measured at time $t + \Delta T$). Thus, we can compare the correlations between these metrics and node's latent degree centrality quantitatively, and we show the p-value associated with each feature and their corresponding significance level in Table 3. These data are presented in terms of histogram (See main text Figure 2 (c)).

We observe (see Table 3) that the centrality measures are not performing well in describing a node's future degree centrality except degree centrality and betweenness centrality (1 star significance level, see Table 3), while several positions are significantly better in describing a node's latent degree centrality. For the *user influential probability* measure [21], the historic information of centrality does not give a promising indication of **IN** and **NIN**. In the experiment the sets of $\mathbf{IN}_{t+\Delta T}$ and $\mathbf{NIN}_{t+\Delta T}$ are labeled by the *degree centrality*, however we notice that *degree centrality* metric does not have a very significant correlation with node's future degree centrality. This implies that the *preferential attachment* is not the only dimension in the social network evolution as stated in [37] [38] [39]. While for the different positions, we have several observations: 1) we unfold that different positions have different ability in describing node's future degree centrality; 2) three of them are much better than centrality measures. To **summarize**, even though these state-of-the-art centrality measures are proved to be good at centrality quantification, they are inherently not powerful enough to depict the node's future degree centrality. Because the preferential attachment is not the only origin underlying the social network evolution. Additionally we can observe that positions in triad structures combines

Table 3: Significance of Features

Features	p-value	significance level
Degree Centrality	0.0583	*
Clustering Coefficient	0.5053	
Closeness Centrality	0.7936	
Betweenness Centrality	0.0937	*
PageRank	0.1423	
User Influential Probability	0.2209	
NPP Position 1	0.7388	
NPP Position 2	0.0385	**
NPP Position 3	$1.059e^{-3}$	***
NPP Position 4	$1.55e^{-4}$	****
NPP Position 5	0.31080	

*: p < 0.1; **: p < 0.05; ***: p < 0.01, ****: p < 0.001.

two principles. Triad position 1 and 4 reflect the effect of preferential attachment, while triad position 3 follows the triadic closure principle. This confirms our propositions made above and provides a possible way to balance the effects between preferential attachment and triadic closure.

As *triadic closure* principle suggests, for the unclosed triad (*triad* 2) new links are formed between nodes in position 3, however we have observed that nodes in position 4 is more likely to be important in future. One possible reason underlying such phenomenon is the *preferential attachment* principle, nodes in position 4 have higher attractiveness of links. However in Table 3 we observe that *degree centrality* does not have a comparable significance as the position 4, this suggests that the *preferential attachment* principle is not the only mechanism underlying this.

To further study this effect, we calculated the conditional probability of position 3 and position 4, $Prob(3|4)$ states the probability that a node shows up in position 3 given the condition that it is located in position 4; $Prob(4|3)$ is the probability that a node is located in position 4 given the condition that it is also in position 3. We can see in Figure 5(c) that nodes in position 4 have extremely high probability to be located in position 3 (close to 1.0), while nodes in position 3 have less than 0.3 probability to occur in position 4. This means, nodes in position 4 are affected by both mechanisms of *preferential attachment* and *triadic closure*, while nodes in position 3 are mainly influenced by the *triadic closure* principle. This explains why position 4 has higher significance level than position 3, and further confirms that the *triadic closure* principle is more significant than the *preferential attachment* in social networks evolution. Also this implies an **important** characteristic of the **NPP** method, the node prominence profile combines two well know social principles (i.e. *preferential attachment* and *triadic closure*).

2.4 Prominence: Centrality and Position

In order to demonstrate that prominence is not only represented in the node's centrality (typically measured by *centrality* metrics) but also in the node's position in local structure, we provide a detailed investigation into their interaction from the perspective of *influence events* and provide the evidence that the **NPP** is able to modeling both centrality and position information. In order to validate their connections, we define *link influence* between two nodes u and v.

Definition 6. For a given node u in the time-varying network $G = (V, E, T_V, T_E)$, u is said to have a **link action** on node w at time t if $(u, w) \in E$ and $t \in T_E(u, w)$. T_V is the log of nodes joining timestamps, while T_E is the log of edge formation timestamps.

Additionally we provide the definition of the *link influence* of node u on its neighbor v as follows:

Table 4: **Degree Centrality Status vs. Link Influence Event**

Patterns	1XX	0XX	X1X	X0X	XX1	XX0	11X	00X	10X	01X
Condmat	1530	365	1513	382	95	1800	1316	168	214	197
DBLP	1377	438	1329	486	15	1800	681	498	369	267
Enron	11769	249	11787	231	187	11831	11549	11	220	238
Facebook	6203	2775	6196	2782	10	8977	4794	1373	1409	1402

Definition 7. A node u is said to have a **link influence** on its neighbor v *iff*: 1) there is a link action of node u with another node w at time t; 2) there exists a link action of node v with node w at time t'; 3) $min(T_E(u,v)) < t < t'$ and $t' - t < \sigma$

The σ is the average action delay between two nodes u and v. An example of *link influence* is presented in Figure 10 (left).

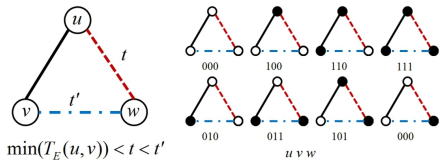


Figure 10: Link Influence Events

We divide the nodes into two groups (important nodes and non important nodes). In this section, we further study the connection between the node's centrality and its position. In Figure 10, we partition the *link influence* event into 8 categories based on nodes' prominence. The three digits represent the degree centrality status of the three nodes, u, v, and w, '1' indicates important node and '0' indicates non important node. In Table 4 we provide the distribution of several patterns, and we observe that: 1) $|1XX| > |0XX|$ and $|X1X| > |X0X|$, this means important nodes have much higher probability to have *link influence* on their neighbors, and it also validates the principle of *preferential attachment*; 2) additionally $|XX0| > |XX1|$, non-important nodes play an important role to transfer *link influence*; 3) $|11X| > |00X|$, this states that *link influence* is more likely to happen between important nodes; 4) $|10X| \approx |01X|$, if *link influence* occurs among important nodes and non-important nodes, then important nodes and non-important nodes have the same chance to initiate the influence. This also validates the interactions between the centrality and position (link formation leads to the change of node's position).

Significance Validation Here we validate the significance of our findings in Table 4. In Table 4 we observed that (i) $|1XX| > |0XX|$ and $|X1X| > |X0X|$; (ii) $|XX0| > |XX1|$; (iii) $|11X| > |00X|$; and (iv) $|10X| \approx |01X|$. In order to validate the significance of these findings, we design the following experiments. In order to demonstrate that these observations can not be explained by degree centrality of nodes alone, we compare the differences (for example, $|1XX| - |0XX|$, or $|XX0| - |XX1|$) to the null model where the degree centrality sequence of network is preserved. The null hypothesis here is that the observed differences in numbers of link influence events can be explained by the degree centrality of the nodes alone, without taking their position in a triad into consideration. To measure statistical significance of these findings, we compare each real network (Condmat, DBLP, Enron, and Facebook) to 10,000 surrogate networks where the degree centrality sequence is preserved and links are placed completely randomly.

In each round of simulation, we generate a random graph G'. In the original network G, the nodes set is denoted as $V = \{v_1, v_2, ..., v_n\}$, which has a corresponding degree centrality sequence $d = \{d_1, d_2, ..., d_n\}$. Realize a random graph G' from the degree centrality sequence d by using

Table 5: **Significance Validation of Link Influence Event By Comparing to Random Graphs with the Same Degree Centrality Sequence (Z-score). Red color indicates the significance level is less than 2 sigmas. (H_0 rejected at 2 sigmas)**

Validation	$\|1XX\| - \|0XX\|$	$\|X1X\| - \|X0X\|$	$\|XX0\| - \|XX1\|$	$\|11X\| - \|00X\|$	$\|10X\| - \|01X\|$
Condmat	7.922	4.937	10.562	7.931	0.975
DBLP	2.350	1.963	4.505	2.023	1.857
Enron	4.258	4.897	4.486	4.583	1.079
Facebook	4.881	4.132	7.691	4.689	1.100

Havel-Hakimi algorithm [47] (in the process of Havel-Hakimi algorithm, we randomly select node in each step). And 8 types of influence events are recalculated from the simulated graph G'.

Here we give an example to show how the validation is performed. For example, in Condmat $|1XX| - |0XX| = 1175$, in order to validate that $|1XX| > |0XX|$ is significant, in each round of simulation, we realize a random graph G' with the same degree centrality sequence as the real network and recalculate the numbers of $|1XX|$ and $|0XX|$ in the simulated network G'. Trivially we have a simulated difference between $|1XX|$ and $|0XX|$, denoted as D_{sim}. After 10,000 round of simulations, statistical significance of the difference between $|1XX|$ and $|0XX|$ is measured by the Z score:

$$Z = \frac{D_{real} - \overline{D}_{sim}}{std(D_{sim})}$$

Clearly in Table 5 we observe that $|1XX|$ is significantly larger than $0XX$ in four real-world networks ($Z >= 2.350$, H_0 rejected). Similarly we also validate that $|X1X| > |X0X|$ and $|11X| > |00X|$ significantly comparing to null model. The observation that $|XX0| > |XX1|$ is significant in four real-world networks (more than 2 sigmas). Finally we also identify that the difference between $|10X|$ and $|01X|$ is not significant in all four real-world networks (less than 2 sigma). This confirms that the difference between $|10X|$ and $|01X|$ is not substantially large. This experiment provides significant evidence for our conclusions in the above section, and demonstrates that our observations can not be explained by the graph's degree centrality sequence alone.

3 Inferring Future Degree Centrality

In order to prove the correctness of our framework, we apply our approach in degree centrality prediction problem and compare with baseline methods. Note that we classified nodes as **IN** or **NIN**, thus making it a binary classification task. We first discuss the feature vector construction aspect.

3.1 Feature Vector Engineering

We first integrate the various measures capturing the notion of centrality in to one feature vector. In addition to the different measures described in Table 3 (other than five positions), we also include some measures introduced in Burt's work of [20], such as *efficiency*, *constraint* and *hierarchy*. These features contribute to the feature vector for the **All** method.

The five TPP positions census contributes to our **NPP** method for prediction. The features for **All** method, **NPP** , **PA** and **TC** are listed in Table 6. For all methods, we use Bagging with *Logistic Regression* as the supervised learning model. Our goal here is to evaluate the utility of additional information imputed by us in the feature vector versus the quality of a learning algorithm.

Table 6: **Features List**

Features	**All**	**NPP**	**PA**	**TC**
Degree Centrality	√		√	
Betweenness	√			
Closeness	√			
Clustering Coef.	√			
PageRank	√			
Efficiency	√			
Hierarchy	√			
Constraint	√			
Position 1		√		
Position 2		√		
Position 3		√		√
Position 4		√		
Position 5		√		

3.2 Experimental Settings

In our experiment we only allow methods to observe features of nodes in a short duration after nodes arriving, for example, for Condmat and DBLP we only use the first year data of new arriving nodes and for Enron and Facebook we only use the first month data of new arriving nodes. We classify the nodes in to **IN** and **NIN** using *degree centrality*.

3.3 Degree Centrality Prediction

3.3.1 Classification Performance

Table 7: Predict Future Degree Centrality. We solve the future degree centrality prediction problem using supervised learning method. The five **NPP** positions (Figure 8) census contributes to our **NPP** method for prediction. **PA** (preferential attachment) method just includes the degree centrality feature, and **TC** (triadic closure) method includes the position 3 census of nodes. **All** method includes existing centrality measures listed in SI. The supervised learning task is to predict whether a new arriving node will become a important node or a non important node (determined by its degree centrality, see SI) in future. **All+** method includes one more feature, $K_{n,n}$, the average nearest neighbor degree centrality. The experiment settings are provided in SI.

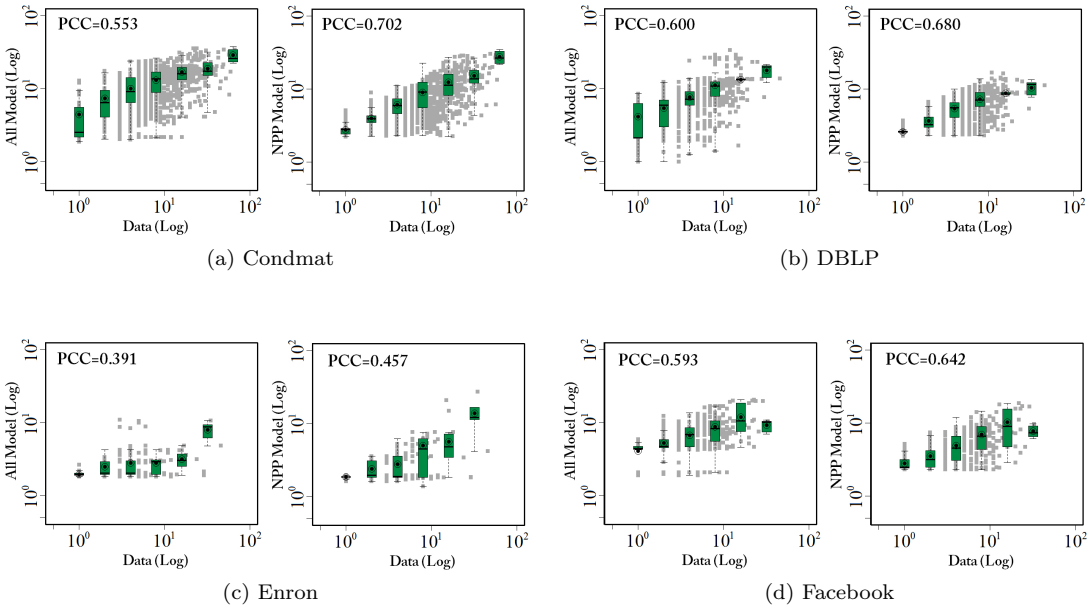
	AUC						AUPR					
Datasets	PA	TC	All	All+	$K_{n,n}$	NPP	PA	TC	All	All+	$K_{n,n}$	NPP
Condmat	0.85	0.72	0.85	0.76	0.71	**0.86**	0.68	0.42	0.71	0.37	0.31	**0.72**
DBLP	0.79	0.83	0.72	0.74	0.73	**0.85**	0.27	0.34	0.19	**0.36**	0.33	**0.36**
Enron	0.71	0.55	0.70	0.51	0.52	**0.72**	0.43	0.18	0.51	0.17	0.15	**0.52**
Facebook	**0.81**	0.78	0.74	0.56	0.74	**0.81**	0.42	0.32	0.42	0.40	0.08	**0.45**

In Table 7, we provide an empirical comparison of learning performance. In our observation our approach **NPP** outperforms the three baseline methods in terms of $AUPR$, and has better or comparable performance in terms of AUC. We have several **conclusions**: 1) the principle *preferential attachment* is just one dimension of mechanisms underlying the nodal degree centrality evolution; 2) the trade-offs between *triadic closure* and *preferential attachment* are well balanced in *node prominence profile* and then it achieves better performance in the prediction task.

Besides we also validate the performance of $K_{n,n}$—the average nearest neighbor degree centrality. In Table 7 we can observe that although $K_{n,n}$ has comparable performance as *node prominence profile* in DBLP dataset, it does not perform consistently well in other datasets. For example, in Facebook, $K_{n,n}$ has much worse performance than others. In all four datasets, our method *node prominence*

profile has better or comparable performance than *All+* method ($K_{n,n}$ is included in the feature vector).

3.3.2 Degree Centrality Prediction Performance



(a) Condmat (b) DBLP

(c) Enron (d) Facebook

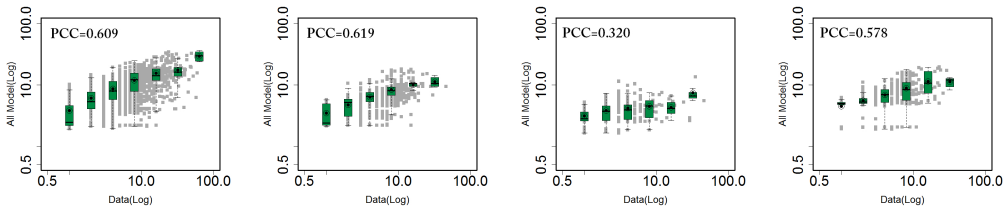
Figure 11: Degree Centrality Prediction Performance. Comparing the measured degree centrality (log scale) with the predicted degree centrality (log scale) in four real-world networks. In each subfigure, the left side is the performance of **All** model and the right side is the performance of **NPP** model. The black concentric circles represent the average predicted values in each data bin. For each data bin we also provide the boxplot of the corresponding predicted values. The performance is measure by the Pearson Correlation Coefficient (PCC). Higher PCC value indicates more accurate prediction of future degree centrality.

Table 8: Predict Degree Centrality. Besides comparing **NPP** method with **All** method, we also compare our **NPP** method with other five state-of-art methods and **All+** (including one more feature than **All** method, $k_{n,n}$) method in terms of Pearson Correlation Coefficient (PCC).

Datasets	Pearson Correlation Coefficient								
	Betweenness	Clustering coef.	Degree Centrality	PageRank	Closeness	$K_{n,n}$	All	All+	NPP
Condmat	0.073	0.378	0.509	0.426	0.162	0.217	0.533	0.609	**0.702**
DBLP	0.216	0.251	0.422	0.113	0.139	0.332	0.600	0.619	**0.680**
Enron	0.308	0.383	0.372	0.228	0.036	0.015	0.391	0.320	**0.457**
Facebook	0.391	0.352	0.561	0.256	0.269	0.285	0.593	0.578	**0.642**

Besides predicting whether new arriving nodes become **IN** or **NIN** in future (classification prediction), our model is also able to predict future degree centrality of these new arriving nodes. The experimental settings and feature vectors for **All** model and **NPP** model are the same as the classification prediction task (in Section 3.3.1). The only difference is, we employ Bagging with *Random Subspace* as the supervised learning model for the degree centrality prediction task.

In Figure 11 we provide the performance of **All** method and **NPP** method in predicting nodes' future degree centrality. The performance of degree prediction is measured by the Pearson Correlation Coefficient (PCC), higher correlation coefficient indicates better performance. In all four



(a) Condmat (b) DBLP (c) Enron (d) Facebook

Figure 12: Degree Centrality Prediction Performance of All+ (including $k_{n,n}$ in feature vector). Comparing the measured degree centrality (log scale) with the predicted degree centrality (log scale) in four real-world networks. The performance is measure by the Pearson Correlation Coefficient (PCC). Higher PCC value indicates more accurate prediction of future degree centrality.

real-world network, **NPP** method outperforms **All** method. The improvement of **NPP** method over **All** method ranges from 8.26% to 26.94%. Additionally we also compare our method with five state-of-art measures (betweenness, degree centrality, clustering coefficient, pagerank, and closeness), results in Table 8 demonstrate that our method **NPP** has the best performance in four real-world networks. The performance of **NPP** method in predicting future degree centrality provides more evidence for our conclusions made above. Our methodology (**NPP**) is validated to optimize trade-offs between essential dimensions of network evolution (*preferential attachment* and *triadic closure*), and yields accurate and generic performance in predicting node's future degree centrality (either classification task or regression task).

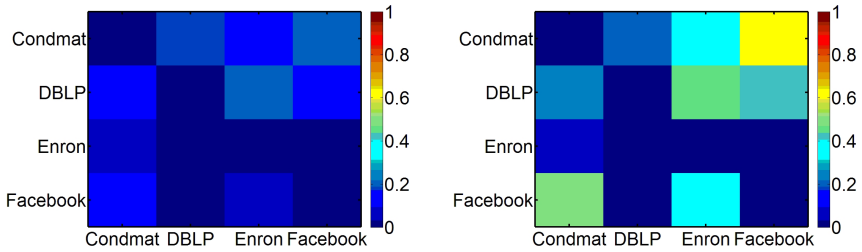
Similar to the above section, we also provide the performance of $K_{n,n}$—the average nearest neighbor degree centrality. In Table 8 we can observe that although $K_{n,n}$ has promising performance in DBLP dataset ($PCC = 0.332$), it does not perform consistently well in other datasets. Additionally including $K_{n,n}$ into the **All** method does not necessarily improve the performance of prediction. For example, in Enron and Facebook **All+** method has worse performance than the **All** method. In all four datasets, our method *node prominence profile* has better performance than $K_{n,n}$ and **All+** method ($K_{n,n}$ is included in the feature vector). Different from the binary degree centrality status prediction task, predicting future degree centrality uncovers the differences between methods in a more detailed way.

4 Generalization across Datasets: A case for transfer learning

In the above sections we have demonstrated that the *node prominence profile* has a stronger generalization capacity than nodal attributes based methods in predicting future important (with high degree centrality) nodes. To be rigorous, we now ask: are these features powerful enough to transfer learning from one social network to another? If our framework are able to generalize across datasets, then it will further demonstrate that our framework captures the essential principles of network evolution.

4.1 Generalization-the Degree Centrality Prediction

We first consider the degree centrality prediction problem. In Figure 13, we provide the transferred learning results for **All** model and **NPP** model. Each pair of generalization is trained on the row dataset and evaluated on the column dataset by Bagging with *logistic regression*. The diagonal entries represent the performance of models which are trained and tested on the same dataset,



(a) NPP (b) All

Figure 13: Generalization Measured in AUPR (Prominence Prediction). Different from the learning task performed in single dataset, the training set is extract from one dataset and the prediction (testing set) is made on another dataset. AUPR, area under precision-recall curve. The AUPR score is more sensitive than AUROC in reflecting the difference of prediction [18]. In order to demonstrate stability of generalization, we use AUPR for the performance evaluation. The detail of **NPP** and **All** methods can be found in **Supporting Information, 4**. Each element represents the performance reduction compared with the regular learning results (i.e., training and testing on the same dataset). The diagonal entries represent the performance of models which are trained and tested on the same dataset, which makes it convenient for comparisons. We can observe that the performance reductions of **NPP** method are mostly less than 20%, while the performance reduction of **All** method can achieve about 60%.

which makes it convenient for comparisons.

There are several observations. We observe that the **NPP** method's performance degrades remarkably less than the **All** method in most cases. This indicates that the prominence profile of node captures principles that are more generic than the centrality based model, and this still holds even if the generalization is across different domains of networks. This further confirms that the prominence profile is a general cross-domain property for the degree centrality evolution analysis. In conclusion, the prominence profile is notably more generic across different domains of networks, and the centrality based method is more particular to a specific dataset.

In conclusion based on the generalization of the degree centrality prediction problem across datasets, we postulate that the positions where nodes are located are more important in determining their evolution orbits than the nodal attributes possessed by them. Our methodology of prominence profile has a greater degree of precision than has heretofore been possible in depicting the network evolution. This is due to the optimized trade-offs between *triadic closure* and *preferential attachment* in our node prominence profile methodology.

References

[1] Kempe, D., Kleinberg, J. & Tardos, E. Maximizing the Spread of Influence through a Social Network. *Proc. 9th SIGKDD*, 137-146 (2003).

[2] Chen, W., Wang, Y. & Yang, S. Efficient Influence Maximization in Social Networks. *Proc. 15th SIGKDD*, 199-208 (2009).

[3] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J. & Glance, N. Cost-effective outbreak detection in networks. *Proc. 13th SIGKDD*, 420-429 (2007).

[4] Scripps, J., Tan, P. N. & Esfahanian, A. H. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. *Proc. 15th SIGKDD*, 747-756 (2009).

[5] Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab*, (1999).

[6] Freeman, L. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35-41 (1977)

[7] Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, 581-603 (1966).

[8] Holland, P. W. & Leinhardt, S. Transitivity in structural models of small groups. *Comparative Group Studies* **2**, 107-124 (1971).

[9] Newman, M. E. J. Power laws, Pareto Distributions, and Zipf's law. *Contemp. Phys.* **46**, 323-351 (2005).

[10] Sharara, H., Singh, L., Getoor, L. & Mann, J. Finding Prominent Actors in Dynamic Affiliation Networks. *Human Journal* (2012).

[11] Dwork, C., Kumary, R., Naorz, M. & Sivakumarx, D. Rank Aggregation Methods for the Web. *Proc. 10th WWW*, 613-622 (2001).

[12] Holland, P. & Leinhardt, S. An Exponential Family of Probability Distributions for Directed Graphs. *JASA* **76**, 33-50 (1981).

[13] Agarwal, N., Liu, H., Tang, L. & Yu P. S. Identifying the Influential Bloggers in A Community. *Proc. 1st WSDM*, 207-218 (2008).

[14] Xiang, R., Neville, J. & Rogati, M. Modeling Relationship Strength in Online Social Networks. *Proc. 19th WWW*, 981-990 (2010).

[15] Tang, L. & Liu, H. Relational Learning via Latent Social Dimensions. *Proc. 15th SIGKDD*, 817-826 (2009).

[16] Scellato, S., Noulas, A. & Mascolo, C. Exploiting place features in link prediction on location-based social networks. *Proc. 17th SIGKDD*, 1046-1054 (2011).

[17] Sun, Y., Han, J., Aggarwal, C. C. & Chawla, N. V. When will it happen? Relationship prediction in heterogeneous information networks. *Proc. 5th WSDM*, 663-672 (2012).

[18] Litchenwalter, R. & Chawla, N. V. Vertex Collocation Profiles: Subgraph Counting for Link Analysis and Prediction. *Proc. 21st WWW*, 1019-1028 (2012).

[19] Barabasi, A. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**, 509-512 (1999).

[20] Burt, R. S. Structural Holes: The Social Structure of Competition. *Havard University Press* (Cambridge, MA, 1992).

[21] Goyal, A., Bonchi, F. & Lakshmanan, L. V. Learning Influence Probabilities in Social Networks. *Proc. 3rd WSDM*, 241-250 (2010).

[22] Grandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. & Suri, S. Feedback Effects between Similarity and Social Influence in Online Communities. *Proc. 14th SIGKDD*, 160-168 (2008).

[23] Katz, E. The two-step flow of communication: an up-to-date report of an hypothesis. *Public Opin. Q.*, 61-78 (1973).

[24] Liben-Nowell, D. & Kleinberg, J. The Link Prediction Problems for Social Networks. *J. Assoc. Inf. Sci. Technol.* **58**, 1019-1031 (2007).

[25] Litchenwalter, R. N., Lussier, J. T. & Chawla, N. V. New Perspectives and Methods in Link Prediction. *Proc. 16th SIGKDD*, 243-252 (2010).

[26] Papadopoulos, F., Kitsak, M., Serrano, M. A., Boguna, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537-540 (2012).

[27] Deng, H., Han, J., Zhao, B., Yu Y. & Lin C. X. Probabilistic topic models with biased propagation on heterogeneous information networks. *Proc. 17th SIGKDD*, 1271-1279 (2011).

[28] Leskovec, J., Lang, K., Dasgupta, A. & Mahoney, M. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Arxiv preprint arXiv: 0810.1355* (2008).

[29] Viswanath, B., Mislove, A., Cha, M. & Gummadi, K. P. On the evolution of user interaction in Facebook. *Proc. 2nd WOSN*, 37-42 (2009).

[30] Litchenwalter, R. & Chawla, N. V. Link prediction: fair and effective evaluation. *Proc. ASONAM*, 376-383 (2012).

[31] Granovetter, M. S. The Strength of Weak Ties. *Am. J. Sociol.* **78**, 1360-1380 (1973).

[32] Klimek, P. & Thurner, S. Triadic Closure Dynamics Drives Scaling-laws in Social Multiplex Networks. *Arxiv preprint arXiv: 1301.0259* (2013).

[33] Szell, M., Lambiotte, R. & Thurner, S. Multirelational Organization of Large-scale Social Networks in an Online World. *Proc. Natl. Acad. Sci* **107**, 13636-13641 (2010).

[34] Heider, F. *The Psychology of Interpersonal Relations* (John Wiley & Sons, 1958).

[35] Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N. V., Rao, J. & Cao, H. Link Prediction and Recommendation across Heterogeneous Social Networks. *Proc. 12th ICDM*, 181-190 (2012).

[36] Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of Growing Random Networks. *Phys. Rev. Lett.* **85**, 4629-4632 (2000).

[37] Leskovec, J., Backstrom, L., Kumar, R. & Tomkins, A. Microscopic Evolution of Social Networks. *Proc. 14th SIGKDD*, 462-470 (2008).

[38] Jin, E. M., Girvan, M. & Newman, M. E. J. Structure of Growing Social Networks. *Phys. Rev. E.* **64(4)**, 046132 (2001).

[39] Li, M., Gao, L., Fan, Y., Wu, J. & Di, Z. Emergence of global preferential attachment from local interaction. *New J. Phys.* **12**, 043029 (2010).

[40] Lou, T., Tang, J., Hopcroft, J., Fang, Z. & Ding, X. Learning to Predict Reciprocity and Triadic Closure. *TKDD* **7** (2013).

[41] Davidsen, J., Ebel, H. & Bornholdt, S. Emergence of a small world from local interactions: Modeling acquaintance networks. *Phys. Rev. Lett.* **88**, 128701 (2002).

[42] Marsili, M., Vega-Redondo, F. & Slanina, F. The rise and fall of a networked society: A formal model. *Proc. Natl. Acad. Sci.* **101**, 1439 (2004).

[43] Toivonen, R., Kovanen, L., Kivela, M., Onnela, J. P., Saramaki, J. & Kaski, K. A comparative study of social network models: Network evolution models and nodal attribute models. *Soc. Networks* **31**, 240 (2009).

[44] Watts, D. J. & Strogatz, S. H. Collective Dynamics of 'Small-World' Networks. *Nature* **393**, (6684):409-10 (1998).

[45] Engle, R. F. Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics *Handbook of Econometrics II*. Intriligator, M. D. and Griliches, Z. (ed.) 796-801 (Elsevier, 1984).

[46] Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).

[47] Hakimi, S. L. On realizability of a set of integers as degrees of the vertices of a linear graph. I. *Journal of the Society for Industrial and Applied Mathematics* **10**, 496-506 (1962).