SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPLEX NETWORKS
COMPUTER SCIENCE

Received
26 June 2014

Accepted
5 November 2014

Published
28 November 2014

Correspondence and requests for materials should be addressed to N.V.C. (nchawla@nd.edu)

Predicting Node Degree Centrality with the Node Prominence Profile

Yang Yang, Yuxiao Dong & Nitesh V. Chawla

Interdisciplinary Center for Network Science and Applications (iCeNSA), Department of Computer Science and Engineering, University of Notre Dame.

Centrality of a node measures its relative importance within a network. There are a number of applications of centrality, including inferring the influence or success of an individual in a social network, and the resulting social network dynamics. While we can compute the centrality of any node in a given network snapshot, a number of applications are also interested in knowing the potential importance of an individual in the future. However, current centrality is not necessarily an effective predictor of future centrality. While there are different measures of centrality, we focus on degree centrality in this paper. We develop a method that reconciles preferential attachment and triadic closure to capture a node's prominence profile. We show that the proposed node prominence profile method is an effective predictor of degree centrality. Notably, our analysis reveals that individuals in the early stage of evolution display a distinctive and robust signature in degree centrality trend, adequately predicted by their prominence profile. We evaluate our work across four real-world social networks. Our findings have important implications for the applications that require prediction of a node's future degree centrality, as well as the study of social network dynamics.

Social networks spurred by digital innovations, such as Facebook, LinkedIn, and Twitter, make up an increasingly wide range of diverse human interactions. These social networks are dynamic and evolve over time, wherein new nodes enter a network, new links may form between nodes or old links may diminish between nodes, and a node's centrality may change over time. Thus, the node and the network co-evolve, where the node impacts the network and the network impacts the node, creating an intertwined effect of centrality and relative position of the node[1]. As the network evolves, we are interested in knowing the predictability of centrality of a node. 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[2,3]).

Over the last decade, network evolution modeling focused on defining basic mechanisms driving link creation and capturing different properties observed in real networks, such as power-law degree distribution, small diameter, and clustering coefficient as a function of node degree centrality[8,10,19]. However, the network evolution drives not only the emergence of macroscopic scaling of social networks but also the microscopic behaviors of individuals. The Barabasi-Albert model[8] provides a mechanism for the emergence of scale-free property in social networks, where new links are established preferentially to well connected individuals. It is also evident that preferential attachment is not sufficient to reproduce other important features of social networks, and individual's link formation also significantly relies on its neighbors[12,20]. The principle of triadic closure has been empirically demonstrated to be relevant for several macroscopic scaling laws in the work of[9,17–23], explicitly or implicitly. The triadic closure mechanism is based on the premise that two individuals with mutual friends have a higher probability to establish a link. These two principles successfully capture the main characteristics of social networks. However preferential attachment requires global information while triadic closure only needs local information[20]. Triadic closure captures the notion of relative position of a node.

Irrespective of the specific mechanisms that act to drive the emergence of macroscopic scaling of social networks, it is reasonable to ask whether such mechanisms also shape the microscopic behaviors of individuals — the degree centrality change. Can we effectively predict degree centrality of a node in the future? We find that the current degree centrality is a weak predictor of the future degree centrality. Rather, the degree centrality evolution is an artifact of both the centrality (preferential attachment) of the node and its relative position (triadic closure) in the network, and is a suitable proxy for interpreting the evolution of a network from a microscopic perspective. We define this combination of centrality and position as prominence. A node may become important over time, which may be a result of its individual achievement or neighborhood structure, which represent aspects of preferential attachment and triadic closure respectively. A node is prominent if the links to the node make it

visible to the other nodes in the network[5,6]. The prominence of a node also depends on the overall structure of its neighborhood. This paper develops a methodological framework that characterizes the prominence of an individual by reconciling the trade-offs between preferential attachment and triadic closure, that is the microscopic level, and develops a model to predict degree centrality of a node in the future. We call our framework the *Node Prominence Profile (NPP)*.

Formally, the *Node Prominence Profile* is defined as follows:

Definition 1. 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 of three isomorphic triad substructures (Figure 1).

In Figure 1 we demonstrate 5 automorphism positions in 3 triad sub-structures. These three triad sub-structures were discussed in social balance theory proposed by Heider[30]. To compute the node prominence profile for an individual node v, we just need to find out all triad sub-structures where node v is located; and then we count how many times node v occurs in each automorphism position (see Supporting Information, 2.3). In this way a high degree centrality node tends to be located in many triad 1 and triad 2 sub-structures, and has high occurrence frequencies in position 1 and position 4 correspondingly. Based on the principle of preferential attachment, they are more likely to attach new links in future. Nodes in position 2 are not necessarily isolated, they just do not have direct links with other two nodes in such a sub-structure (Triad 1). Clearly nodes in position 2 have potential to develop new links in future. Driven by triadic closure effect, triad 2 tends to evolve to triad 3, thus nodes having high occurrence frequencies in position 3 are more likely to attach new links. Triad 3 is a stable sub-structure[30], as the propensity of attaching new links for the nodes having high occurrence frequencies in position 5 is relatively small. Positions in these three triad structures embody both principles — preferential attachment and triadic closure.

The empirical experiments reveal that **NPP** is able to provide more precise prediction of node's future degree centrality over baseline solutions. **NPP** is validated on four different social networks. We also demonstrate that the model developed on one social network and predict on another social network (transfer learning), thus demonstrating the generalization capacity of **NPP** and confirming that it is effective in capturing the general factors underlying social network evolution impacting the degree centrality of a node.

Results

Node Prominence Profile. As we posited, prominence is not only represented in the node's centrality but also in the node's position in local structure. Much effort has been devoted to measuring the node's centrality, such as degree centrality, Pagerank[4], Betweenness[24], and Closeness[25]. Here we will demonstrate that modeling prominence can lead to a much improved prediction about a node's future degree centrality in the network than modeling current state-of-the-art centrality measures.

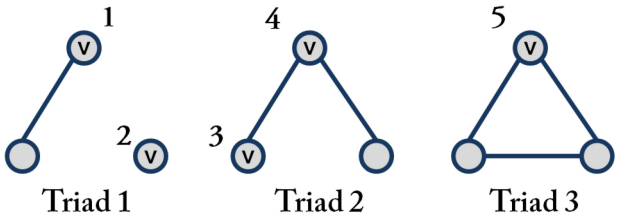

Figure 1 | Node Prominence Profile. In this figure, we mark 5 automorphism positions (labeled with v) in 3 triad structures. The node prominence profile of a node v, is a vector describing the occurrence frequencies of node v in these 5 automorphism positions.

Triadic Closure Effect on Degree Centrality Evolution. The effect of preferential attachment on the degree centrality evolution is evident[8]. However, the prominence of a node not only includes its centrality but also its position in local neighborhood, and the principle of *preferential attachment* is inherently unable to describe node's position in local structure[20]. The triadic closure principle provides us an alternate solution. We first study the effect of triadic closure on the degree centrality evolution. The quantity of triadic closure (or structural balance) can be defined as below[12] (Supporting Information, 2.2):

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

By studying the sub-networks of important (high degree centrality) or non-important (low degree centrality) nodes (based on Pareto Principle[11], we partition nodes into important and non-important nodes based on their degree centrality, denoted as **IN** and **NIN**, see Methods and Supporting Information, 1.2), we observe that initially the sub-network of future important (having high degree centrality in future) nodes has a lower balance rate than the sub-network of future non-important nodes, but the former sub-network evolves to form a more balanced topology (Figure 2 (a)). There are several implications: 1) there exist connections between the triadic closure and the degree centrality evolution. In addition, new links are more likely to form between nodes located in an unbalanced sub-network; 2) The initial sub-network where future important nodes are located is more imbalanced than that of future non-important nodes, thus position of node can be indicative of its future degree centrality. These findings are demonstrated to be statistically significant at 95% confidence even if we scale the threshold value of important nodes, such as 10%, 30% and 50% (see Supporting Information, 2.2). This implies the possible effect of triadic closure on both the node's degree centrality and its position in local neighborhood.

As suggested in the principle of triadic closure, a "forbidden" triad[12] (*triad 2* in Figure 1) is more likely to attach new links. In order to demonstrate that position is crucial for the degree centrality evolution, we provide the evolution ratio of two types of triads in Figure 1. For a triad structure, if there are new links attached, then we say this triad structure evolves. And for a specific type of triads (i.e., triad 1), we calculate how many percentage of them evolve and denote that as the evolution ratio (see Supporting Information, 2.2). We can see that the "forbidden" triad (*triad 2*) has much higher probability to attach a new link than a disconnected sub-structure *triad 1* (Figure 2 (b)). This implies that nodes in different genres of triads have different probabilities to develop degree centrality. This leads us to an important conclusion: the positions of nodes in sub-structures determine their future orbits in both essential prominence elements for describing degree centrality evolution. Our methodological framework called, the Node Prominence Profile (Definition 1), incorporates these insights in the modeling for the node's prominence (Figure 1; (Supporting Information, 2.3)).

Positions in Triad Structure. In Figure 1 we enumerate all possible five positions in the triad sub-structures described in Definition 1. We observe that the position of a node within its local structure is related with its degree centrality evolution. For instance in Figure 2 (b) nodes in "forbidden" triad are more likely to attract new links (Supporting Information, 2.3). Thus, different positions of a node in corresponding triads should have distinct descriptive power of degree centrality evolution.

In Figure 2 (c) we provide the significance of different centrality measures and position incidence values in indicating node's future degree centrality (**IN** or **NIN** in future).

We observe (see Figure 2 (c)) that the centrality measures are not performing well in describing a node's future degree centrality except degree centrality and betweenness centrality. At the same time

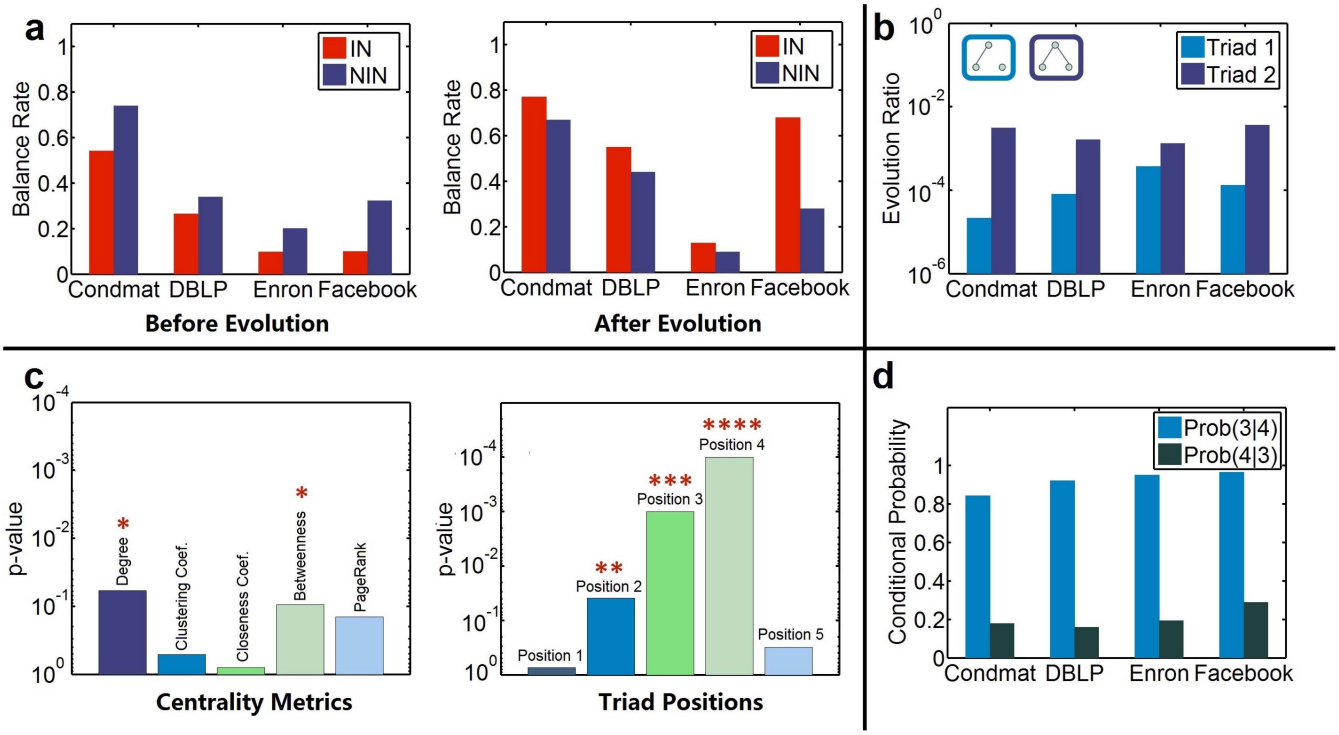

Figure 2 | Microscopic Prominence Analysis. (a) Structural Balance Rate. For the nodes joining the network G_t at the same time t, based on their degree centrality in the network $G_{t+\Delta t}$ after Δt timestamps, we divide them into two sets **Important Nodes** and **Non Important Nodes** (see Supporting Information, 1.2 for detail). (left) In the network G_t we extract the sub-networks of **IN** and **NIN** and calculate their balance rates correspondingly. We observe that the **IN** sub-network has a lower balance rate than the **NIN** sub-network. (right) Similarly in the network $G_{t+\Delta t}$ we extract the sub-networks of **IN** and **NIN**, the **IN** sub-network has a larger balance rate than the **NIN** sub-network. (b) Triad Evolution Rate. In four datasets we compute the link formation probability within different kinds of triads, we call it triad evolution rate. We observe that the "forbidden" triad (*triad 2*) (Figure 1) has much higher probability to form a new link than the disconnected sub-structure *triad 1*.(Supporting Information, 2.3). (c) Significance of Inferring Future Degree Centrality. We consider these centrality measures and positions as predictors of future degree centrality, we show the *p*-value associated with each feature and its corresponding significance level under Wald test (Supporting Information, 2.3). (d) Position Conditional Probability. We calculated the conditional probability of position 3 and position 4 (see Figure 1), *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 that nodes in position 4 have high probability to be located in position 3, while nodes in position 3 have less than 0.3 probability to occur in position 4. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$, ****: $p < 0.001$.

several position incidence values are significantly better in inferring a node's latent degree centrality. In the experiment the sets of $\mathbf{IN}_{t+\Delta T}$ (nodes with high degree centrality at time $t + \Delta T$) and $\mathbf{NIN}_{t+\Delta T}$ (nodes with low degree centrality at time $t + \Delta T$) (Supporting Information, 1.2) are labeled based on the degree centrality (Figure 2). We note that the *degree centrality* metric itself does not have the most significant correlation with node's future degree centrality when ΔT is large (Supporting Information, 1.2). Based on these observations, we have several conclusions: 1) different positions have different power in describing node's future degree centrality; 2) three of them are much better than centrality measures themselves. To **summarize**, even though the centrality measures are demonstrated to be good at centrality (relative importance in the network) quantification, they are not powerful enough to depict the node's future degree centrality. This is because, the preferential attachment is not the only origin underlying the social network dynamics[12,19,20]. Additionally we can observe that positions in triad structures embody both principles—preferential attachment and triadic closure. Triad position 1 and 4 reflect the effect of preferential attachment, while triad position 3 manifests 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 and model two essential elements of node's degree centrality effectively.

As *triadic closure* principle suggests, for the unclosed triad (*triad* 2) new links are attached between nodes in position 3, however we have observed that nodes having high occurrences in position 4 are more likely to have high degree centrality in future. One possible reason underlying such phenomenon is, nodes in position 4 have higher attraction to links (*preferential attachment*). However in Figure 2 (c) we already identify that *degree centrality* does not have a comparable performance as the position 4. To further investigate this, 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 in one triad given the condition that it is located in position 4 in a different triad; *Prob*(4|3) is the probability that a node is located in position 4 given the condition that it is also in position 3. In Figure 2 (d) we can see 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 be in position 4. This means, nodes in position 4 are influenced by both mechanisms of *preferential attachment* and *triadic closure*, while nodes in position 3 are mainly affected by the *triadic closure* principle. This explains why position 4 has higher significance level than position 3. This also reflects an **important** property of these positions in triad structures—they combine the two well known social principles (i.e. *preferential attachment* and *triadic closure*).

Prominence: Centrality and Position. In order to demonstrate that prominence is not only represented in the node's centrality 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 model both centrality and position information. In order to validate their connections, we define *link influence* (Figure 3) between two nodes u and v (Supporting Information, 2.4).

Definition 2. For a given node u in the time-varying network $G = (V, E, T_V, T_E)$ (see Supporting Information, 1.2 Definition 3), 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:

Definition 3. 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 3.

We divide the nodes into two groups (important nodes and non important nodes), as considered in Figure 2. As shown, in Figure 3, we partition the *link influence* event into $2^3 = 8$ categories based on nodes' degree centrality; '1' indicates an important node (high degree centrality) and '0' indicates a non important node (low degree centrality). In Figure 4 we provide the distribution of several patterns, and we observe that: 1) $|1XX| > |0XX|$ and $|X1X| > |X0X|$ ($|1XX|$ is the number of link influence events where node u is an important node, Figure 3), 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. These patterns persist in four different real-world networks and are proved to be statistically significant (see Supporting Information, 2.4). Thus, this further implies that interactions between degree centrality and position (link formation leads to the change of node's position) are common in social networks.

Prediction of Future Degree Centrality. The **NPP** method described above allows us to investigate two aspects of node's prominence. We present empirical analysis to validate that our method is able to more effectively predict a node's future degree centrality than the state-of-art methods.

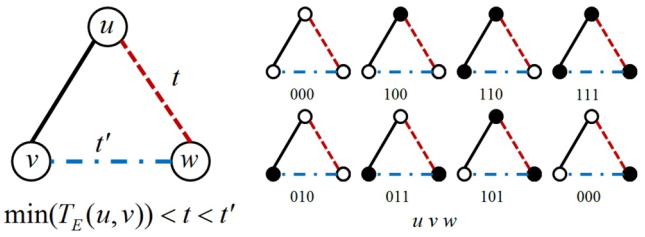

Figure 3 | Link Influence Events. On the left side, we demonstrate the link influence of node u on its neighbor v. Node u has a link action with node w at time t, and node v has a link action with node w at time t'. And the gap between t and t' is smaller than a threshold σ. σ is the average action delay between two nodes u and v. On the right side, we enumerate all possible kinds of link influence events if nodes in the network are partitioned into important nodes and non-important nodes. The three digits encode the degree centrality status of the three nodes, u, v, and w, '1' indicates an important node (high degree centrality) and '0' indicates a non important node (low degree centrality). Thus there are 8 kinds of link influence events.

Inferring Future Degree Centrality. In Table 1, we provide an empirical comparison of the performance for the problem of degree centrality prediction in terms of AUROC (Area under the ROC curve) and AUPR (Area Under the Precision-Recall Curve) (see Methods). Our approach, **NPP**, outperforms three baseline methods in terms of *AUPR*, and has better or comparable performance in terms of *AUROC*. **All** method includes existing state-of-the-art centrality measures, which is described in Supporting Information, 3.1. 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 **NPP** and then it achieves better performance in the degree centrality prediction task. Additionally our model NPP is also able to predict nodes' future degree centrality and yield better performance than the state-of-art methods (see Supporting Information, 3.3). This further confirms the correctness and effectiveness of our methodology.

Generality of Node Prominence Profile: Prediction Across Datasets. We have demonstrated that the **NPP** has a stronger generalization capacity than state-of-the-art centrality measures in predicting future degree centrality. To be rigorous, we now ask: are these features powerful enough to transfer learning from one social network to another? That is, can a model developed on one social network effectively predict for another social network? If our framework is able to generalize across datasets, then it will further demonstrate that our framework captures the essential principles of degree centrality evolution.

Generalization of the Degree Centrality Prediction. In Figure 5, we provide the transfer learning (transfer of learning is usually described as the process and the effective extent to which past experiences (trained model) affect performance (prediction) in a new situation or data different from the one that the model was trained on) results for **All** model and **NPP** model. Each pair of generalization is trained on the row dataset and evaluated on the column dataset. Transfer learning generally leads to the loss in performance. In Figure 5 we provide the performance loss of transferred learning compared with the performance of non-transfer learning (where training and testing are conducted on the same dataset). Thus, the diagonal entries all have performance loss as zero.

There are several observations. We observe that the **NPP**'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 state-of-the-art centrality measures, 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 state-of-the-art centrality based method is more particular to a specific dataset.

Discussion

We analyzed several principles/mechanisms underlying the network evolution, mainly focusing on two essential elements of the node's prominence: centrality and position. We demonstrated that the position of a node in a local structure is strongly indicative of the degree centrality progression in the social network. Building on this observation, we developed a prediction method referred to as **NPP**. We empirically demonstrated the effectiveness of **NPP** by demonstrating improvement in performance over the state-of-art methods for the problem of degree centrality prediction in four different datasets. We further established the generalization capacity of our methods under a transfer learning scenario — we learned the classifier on one social network (using the proposed features) and tested on another social network. The performance trends clearly showed that our approach

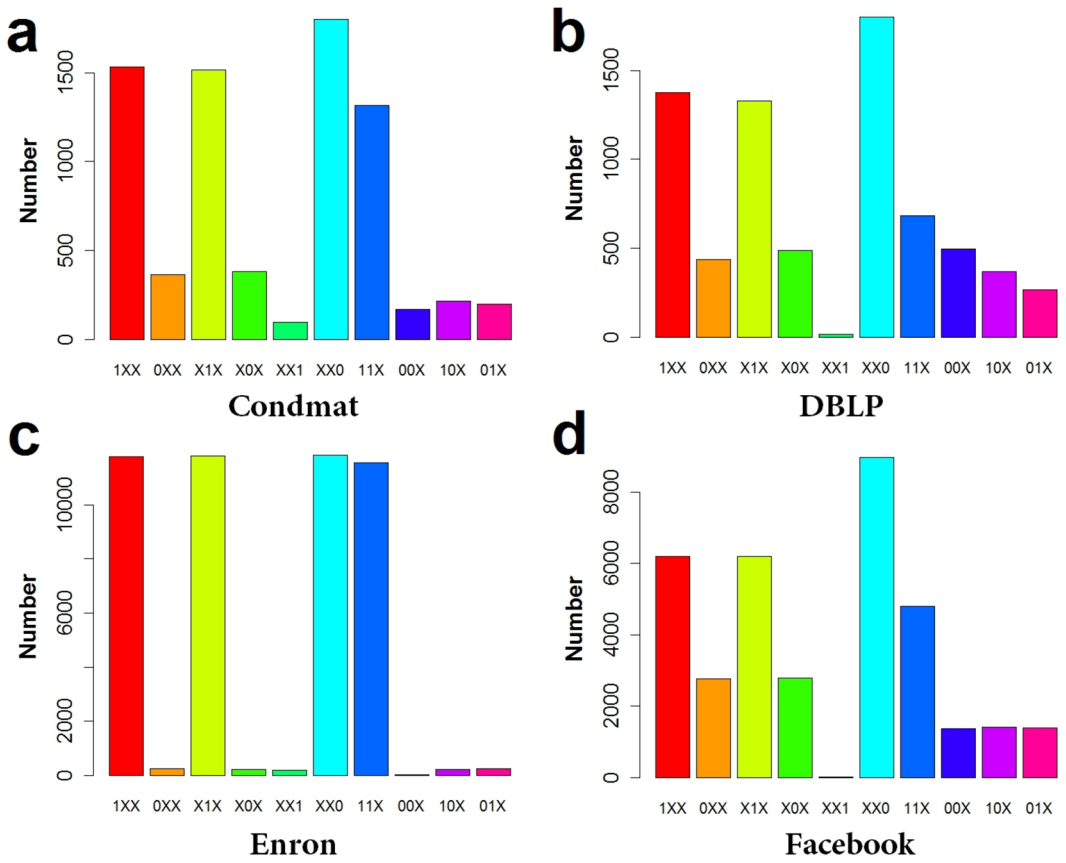

Figure 4 | Degree Centrality Status vs. Link Influence Event. '1' indicates an important node (high degree centrality), '0' indicates a non important node (low degree centrality), and 'X' indicates that the code is either '1' or '0'. We observe similar patterns in four different real-world social networks.

is able to capture essential properties or features underlying degree centrality evolution, which are general across different domains of social networks.

Our methodology (**NPP**) is validated to optimize trade-offs between essential dimensions of network evolution (*preferential attachment* and *triadic closure*). Therefore, it is not surprising that, as a consequence, our approach yields accurate and generic performance in predicting node's future degree centrality. Our method can be effectively used in a variety of applications that rely on inferring a node's importance in the future, as captured by centrality measures. In summary, we have developed a new perspective for growth in degree centrality of a node in a social network and developed a general purpose feature vector that can be used by different machine learning algorithms across different social networks.

Table 1 | Predict Future Degree Centrality. We solve the future degree centrality prediction problem (Supporting Information, 3) using supervised learning method. The five NPP positions (Figure 1) 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 as feature. The method labeled All includes existing centrality measures (Supporting Information, 3.1). 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 Supporting Information, 3.2) in future. The experiment settings are provided in Supporting Information, 3.2

	AUROC				AUPR			
Datasets	PA	TC	All	NPP	PA	TC	All	NPP
Condmat	0.85	0.72	0.85	**0.86**	0.68	0.42	0.71	**0.72**
DBLP	0.79	0.83	0.72	**0.85**	0.27	0.34	0.19	**0.36**
Enron	0.71	0.55	0.70	**0.72**	0.43	0.18	0.51	**0.52**
Facebook	**0.81**	0.78	0.74	**0.81**	0.42	0.32	0.42	**0.45**

Methods

Data Description. In this paper we examine our approaches and perform our analysis on four social networks. The *Condmat* network[13] 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[14] we attach timestamps for each collaboration and choose 3,215 authors who published at least 5 papers. *Enron* dataset[15] 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.[16], which contains wall-to-wall post relationship among 11,470 users between 2004-10 and 2009-01.

Important Nodes and Non Important Nodes. On a global level, important nodes 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 Supporting Information, 1 we observe that top 20% (ranked by PageRank) nodes occupy about 80% *PageRank* centrality in DBLP network. This satisfies *Pareto Principle* (also known as 80-20 rule)[11]. To better understand and model the effects of network evolution on node's prominence, we partition nodes into two sets *important nodes* and *non-important nodes*. **Important Node**: In a network $G = (V,E)$ a node v is a *important node* under centrality measurement $\mathbb{M}$ if and only if $\frac{|\{u|\mathbb{M}(u) \leq \mathbb{M}(v)\}|}{|V|} \geq 0.8$. **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**.

Evaluation Methods. We employ AUROC and AUPR to evaluate the performance of the predictions tasks in this work. The information of associated evaluations metrics are as below:

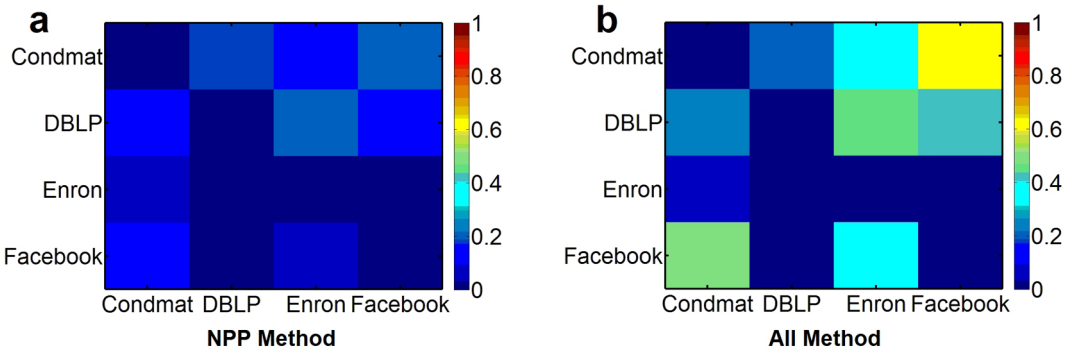

Figure 5 | Generalization Performance Loss in AUPR (Degree Centrality Prediction). Different from the learning task performed in single dataset, the training set is extracted from one dataset and the prediction (testing set) is made on another dataset. AUPR, area under precision-recall curve. The AUPR score is more sensitive than AUROC in reflecting the difference of prediction[7]. 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, 3. **All** method includes existing centrality measures, which is described in Supporting Information, 3.1. Each element represents the performance reduction compared with the regular learning results (i.e., training and testing on the same dataset). 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%.

ROC: The receiver operating characteristic (ROC) represents the performance trade-off between true positives and false positives at different decision boundary thresholds[26,27].

AUROC: Area under the ROC curve.

Precision-recall Curve: Precision-recall curves are also threshold curves. Each point corresponds to a different score threshold with a different precision and recall value[28].

AUPR: Area under the precision-recall curve.

Prediction Experiment Settings. For the prediction of future degree centrality, we use Bagging with *Logistic Regression* as the supervised learning model. Bagging[29] is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. The bagging method reduces variance and helps to avoid over-fitting, which is usually applied to many types of machine learning methods. 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. In our experiment we only allow methods to observe features of nodes in a short duration after nodes joining the network, 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* (Supporting Information, 3).

In order to demonstrate the generality of our framework, we perform the transfer learning for the Degree Centrality Prediction problem. Each pair of generalization is trained on the one dataset and evaluated on another dataset by Bagging with *logistic regression* (for example, trained on Condmat and evaluated on DBLP) (Supporting Information, 4).

1. Gross, T. & Blasius, B. Adaptive coevolutionary networks: a review. *J. R. Soc.* **5**, 259–271 (2008).
2. Rothenberg, R. *et al.* Choosing a centrality measure: Epidemiologic correlates in the Colorado Springs study of social networks. *Soc. Networks* **17**, 273–297 (1995).
3. Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. How Correlated Are Network Centrality Measures? *Connections* **28**, 16–26 (2008).
4. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* **30**, 107–117 (1998).
5. Sharara, H., Singh, L., Getoor, L. & Mann, J. Finding Prominent Actors in Dynamic Affiliation Networks. *Human Journal* **1**, 1–14 (2012).
6. Burt, R. S. & Minor, M. J. (eds.) *Applied Network Analysis: a methodological introduction* (Sage Publications, Newbury Park, 1983).
7. Litchenwalter, R. & Chawla, N. V. Vertex Collocation Profiles: Subgraph Counting for Link Analysis and Prediction. *Proceedings of the 21st international conference on World Wide Web,* Lyon 1019–1028. New York, NY, USA: ACM. (2012 April).
8. Barabasi, A. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).
9. Li, M. *et al.* A coevolving model based on preferential triadic closure for social media networks. *Sci. Rep.* **3**, 2512 (2013).
10. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
11. Newman, M. E. J. Power laws, Pareto Distributions, and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
12. Jin, E. M., Girvan, M. & Newman, M. E. J. Structure of Growing Social Networks. *Phys. Rev. E.* **64**, 4 (2001).
13. Litchenwalter, R. N., Lussier, J. T. & Chawla, N. V. New Perspectives and Methods in Link Prediction. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining,* Washington DC 243–252. New York, NY, USA: ACM. (2010 July).
14. Deng, H. *et al.* Probabilistic topic models with biased propagation on heterogeneous information networks. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego* 1271–1279. New York, NY, USA: ACM. (2011 August).
15. Leskovec, J., Lang, K., Dasgupta, A. & Mahoney, M. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics* **6**, 29–123 (2009).
16. Viswanath, B., Mislove, A., Cha, M. & Gummadi, K. P. On the evolution of user interaction in Facebook. *Proceedings of the 2nd ACM workshop on Online social networks,* Barcelona 37–42. New York, NY, USA: ACM. (2009 August).
17. Klimek, P. & Thurner, S. Triadic Closure Dynamics Drives Scaling-laws in Social Multiplex Networks. *New J. Phys.* **15**, 063008 (2013).
18. 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).
19. Leskovec, J., Backstrom, L., Kumar, R. & Tomkins, A. Microscopic Evolution of Social Networks. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las* Vegas 462–470. New York, NY, USA: ACM. (2008 August).
20. Li, M. *et al.* Emergence of global preferential attachment from local interaction. *New J. Phys.* **12**, 043029 (2010).
21. Davidsen, J., Ebel, H. & Bornholdt, S. Emergence of a small world from local interactions: Modeling acquaintance networks. *Phys. Rev. Lett.* **88**, 28701 (2002).
22. 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).
23. Toivonen, R. *et al.* A comparative study of social network models: Network evolution models and nodal attribute models. *Soc. Networks* **31**, 240 (2009).
24. Freeman, L. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41 (1977).
25. Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, 581–603 (1966).
26. Mason, S. J. & Graham, N. E. Areas Beneath the Relative Operating Characteristics (roc) and Relative Operating Levels (rol) Curves: Statistical Significance and Interpretation. *Q. J. Roy. Meteor. Soc.* **128**, 2145–2166 (2002).
27. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. *ReCALL* **31**, 1–38 (2004).
28. Davis, J. & Goadrich, M. The Relationship Between Precision-recall and ROC Curves. *Proceedings of the 23rd international conference on Machine learning, Pittsburgh* 233–240. New York, NY, USA: ACM. (2006 June).
29. Sahu, A., Runger, G. & Apley, D. Image de-noising with a multi-phase kernel principal component approach and an ensemble version. *Applied Imagery Pattern Recognition Workshop (AIPR), Washington DC* 1–7. New York, NY, USA: IEEE. (2011 October).
30. Heider, F. *The Psychology of Interpersonal Relations* (John Wiley & Sons, New York, 1958).

Acknowledgments

Research was sponsored by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, and by the grant FA9550-12-1-0405 from the U.S. Air Force

Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA).

Author contributions

Y.Y. and N.V.C. designed the research. Y.Y. and Y.D. contributed analytic tools and performed empirical evaluation. Y.Y., Y.D. and N.V.C. analyzed the results. Y.Y.,Y.D. and N.V.C. wrote the paper. The authors declare no conflict of interest. N.V.C. is the corresponding author: nchawla@nd.edu.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Yang, Y., Dong, Y. & Chawla, N.V. Predicting Node Degree Centrality with the Node Prominence Profile. *Sci. Rep.* **4**, 7236; DOI:10.1038/srep07236 (2014).