A time aware method for predicting dull nodes and links in evolving networks for data cleaning

Niladri Sett, Subhrendu Chattopadhyay, Sanasam Ranbir Singh, Sukumar Nandi
Department of Computer Science and Engineering
Indian Institute of Technology, Guwahati
Guwahati, India-781039
Email: {niladri,subhrendu,ranbir,sukumar}@iitg.ernet.in

***Abstract*—Existing studies on evolution of social network largely focus on addition of new nodes and links in the network. However, as network evolves, existing relationships degrade and break down, and some nodes go to hibernation or decide not to participate in any kind of activities in the network where it belongs. Such nodes and links, which we refer as "dull", may affect analysis and prediction tasks in networks. This paper formally defines the problem of predicting dull nodes and links at an early stage, and proposes a novel time aware method to solve it. Pruning of such nodes and links is framed as "network data cleaning" task. As the definitions of dull node and link are non-trivial and subjective, a novel scheme to label such nodes and links is also proposed here. Experimental results on two real network datasets demonstrate that the proposed method accurately predicts potential dull nodes and links. This paper further experimentally validates the need for data cleaning by investigating its effect on the well-known "link prediction" problem.**

Keywords*—*Social network analysis, Dull node, Dull link, Time-series, Link prediction.

I. Introduction

Evolution of social network [1], [2] has been receiving substantial attention in the field of social network analysis (SNA). A social *tie* (or a link) is built through social interaction between two *actors* (or nodes); and structure of a social network evolves with addition of new nodes and links in the network. Association or friendship between two actors is maintained by the amount of information that flows between them, which is often quantified using the pattern of interactions they are involved in. Most of the studies on the evolution of social network have concentrated on addition of new links and nodes in it. However, in reality, some links become inactive with time as friendships break down, and some nodes withdraw themselves from the network. Inaction or disappearance of those links and nodes alter the structure of the network. We refer such nodes and links as *dull nodes* and *dull links* respectively. Ignoring dull nodes and links may affect analysis and prediction tasks in the network, because these nodes and links provide spurious information. This paper proposes a novel time aware method to predict dull nodes and links at an early stage as a *preprocessing* task in social networks. We formally define the task of predicting such nodes and links at an early stage by exploiting their historical properties: `by observing a time-evolving social network up to time` τ`, predicting the nodes (and links) which will be declared dull at a future time` τ'. The nodes and links, which are idle for long, are labeled as dull. A novel scheme is proposed to fix two time duration thresholds, each for nodes and links of a particular network. If a node or link remains idle longer than its respective threshold, it will be labeled as dull. Removal of predicted dull nodes and links from the network reduces noisy information, and is considered as a *data cleaning* step for dynamic networks.

In recent times, researchers have started exploring temporal dimension in SNA. There are studies [3], [4], [5], [6] on two major SNA tasks namely *link prediction* [7] and *community detection* [8], which consider temporal information as an enhancement to baseline methods. These temporal methods inherently carry properties that can discriminate between an active node (or an active link) and a dull node (or a dull link). However, to the best of our knowledge, none of them has formulated and solved the problem of predicting dull nodes (and dull links), nor considered the time-aware data cleaning for dynamic networks. After pruning the predicted dull nodes and links from the network, the preprocessed network can be used for any SNA task using simple non-temporal baseline methods, which may achieve considerable performance gain along with improvement in running time.

Experiments on two real network datasets demonstrate that the proposed method effectively predicts potential dull nodes and links. We further validate our claim of noise reduction by investigating link prediction [1] performance, subject to data cleaning. We show that, the removal of predicted dull nodes and links results in considerable improvement in performance of state-of-the-art link prediction methods for both datasets.

More specifically, contributions of this paper are as follows.

- It proposes a novel data cleaning method for dynamic networks. This method predicts and removes dull nodes and links, which will eventually become inactive or leave the network in future.
- It also proposes a novel scheme to label the dull nodes and links.
- A case study is presented, which demonstrates the effect of the proposed data cleaning method on state-of-the-art link prediction methods.
- Experiments are performed on two real network datasets, which show that the proposed data cleaning

[1] The link prediction problem in social network is defined as: given the snapshot of a network at time t, predicting the links that will appear at a future time t' [7].

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.49

method effectively predicts the dull nodes and links, enhancing link prediction performances.

II. Related works

A brief literature review on existing preprocessing techniques for social networks is presented next. Zhang *et. al.* [9] has proposed SocConnect, which integrates social network data collected from multiple on-line social networking sites. Network data cleaning has been considered in [10], [11], [12], [13]. Benevenuto *et. al.* [10] has proposed a method for detecting spammers in Twitter. Bhagat *et. al.* [11] has developed algorithms for anonymizing the actors in social networks in order to preserve privacy. Ferreira *et. al.* [12] has compiled a nice survey on name disambiguation methods for bibliographic citation networks. Huisman *et. al.* [13] has proposed methods to fill the missing informations in network data. Hernández *et. al.* [14] has proposed algorithms to store and retrieve large social network data in compressed format. Macskassy [15] has proposed a method for identifying nodes which leave a particular community in dynamic networks. To the best of our knowledge, none of the existing studies has attempted to predict dull nodes and links in evolving networks as a network preprocessing task.

We have found a few studies [16], [17], [18], which passively relate to the concept of dull nodes or links. Kamath *et. al.* [16] has proposed a method to track short-term group formation in on-line social networks like Twitter and Facebook. Raeder *et. al.* [17] and Miritello *et. al.* [18] have predicted "persistence" of links. Given the communication pattern of a link in a time window, their methods predict whether the link will remain active in the next window.

III. Problem formulation

This section formulates the problem of predicting dull nodes and links. Social networks typically evolve with social interactions among the actors. When two actors interact, they become connected by a link. With time, pair(s) of actors connected by a link, may interact again. We preserve this history of interaction between two actors by storing the time-stamp of occurrence of the interactions they have been involved in, and associate it to the link formed by them. We represent evolving networks up to time instant t by an undirected graph $G^t = (V, E, T, \mathcal{Q})$, where V is the set of vertices representing actor nodes; $E \subset V \times V$ is the set of edges representing links; T is the set of time-stamps of interactions occurred in the network till t; $\mathcal{Q}$ is a function $\mathcal{Q} : E \rightarrow 2^T$, which returns the historical time-stamps of all interactions associated with a link. For an example, let G^t be a Facebook wall-post network. When a user x posts on another user y's Facebook wall, they become connected by a link (x, y). All such historical wall-posts (up to time instant t) between x and y are stored and retrieved by $\mathcal{Q}(x, y)$.

Given a graph G^t, the preprocessing task is to predict the dull nodes and links, which are going to be dull at a future time t', and construct another graph G^t_r by removing these dull nodes and links from G^t.

IV. Predicting dull nodes and links

Historical activities of nodes and links are modeled to predict the dull nodes and links. First, several time-series representing historical change in baseline node and link properties are prepared and modeled using *simple exponential smoothing* [19] model. Future values of these time-series are forecast to identify features for predicting dull nodes and links. Lastly, vector distance based unsupervised method is applied to predict dull nodes and links.

A. Preparing time-series

The time-series which we are going to introduce in this subsection, are of two types: *dyadic* and *topological*. Dyadic time-series encapsulate the temporal characteristics relating *dyads, i.e.*, the historical interaction between nodes, whereas topological time-series deal with change in graph's topological measures like degree of a node, number of common neighbors of a pair of nodes etc. Given a graph G^t, time is discretized into a sequence of contiguous time-windows of constant size Δ for a particular network. A time-window ending at time $t - \Delta i$ is denoted as $w_{-i} : i \in \mathbb{N}$. All the time-stamps that fall in $\big(t - \Delta(i+1), t - \Delta i\big]$ define the window w_{-i}. w_0 represents the most recent window, and w_{-i} represents the $(i+1)^{th}$ last window. We populate the contiguous time-windows with some value $\in \mathbb{R}$ to form a time-series. The method of populating the time-windows of dyadic and topological time-series is described next.

The window $w_{-\delta}$ of a dyadic time-series $S(x, y)$, defined on a link (x, y) in G^t, is populated as:

$$S_{-\delta}(x, y) = |\mathcal{Q}(x, y) \cap w_{-\delta}|, \tag{1}$$

which gives the number of interactions between the end nodes during time-window $w_{-\delta}$. $S(x, y) = \big\langle S_{-q}(x, y), ..., S_{-1}(x, y), S_0(x, y)\big\rangle$, $q \in \mathbb{N}$, gives the dyadic time-series of the link (x, y), where w_{-q} is the time-window when the link appeared.

Similarly, the window $w_{-\delta}$ of a dyadic time-series $s(x)$, defined on a node x in G^t, is populated as the number of interactions between node x and its neighbors $\Gamma(x)$ in G^t during time-window $w_{-\delta}$:

$$s_{-\delta}(x) = \sum_{y \in \Gamma(x)} |\mathcal{Q}(x, y) \cap w_{-\delta}|, \tag{2}$$

To construct topological time-series of nodes, we consider two node properties namely *degree* and *clustering coefficient (CC)* [2]. We define time-series data for $degree$ and CC corresponding to node x in a time-window $w_{-\delta}$ as follows:

$$d_{-\delta}(x) = d^{t-\Delta\delta}(x) - d^{t-\Delta(\delta+1)}(x), \tag{3}$$

$$c_{-\delta}(x) = c^{t-\Delta\delta}(x) - c^{t-\Delta(\delta+1)}(x), \tag{4}$$

where $d^{t-\Delta\delta}(x)$ and $c^{t-\Delta\delta}(x)$ give the $degree$ and CC of node x respectively in the snapshot of G^t at $t - \Delta\delta$, $G^{t-\Delta\delta}$. We denote these two time-series as $d(x)$ and $c(x)$ respectively. Common neighbor (CN) index [7] is the baseline property

[2] CC represents the local density in the neighborhood of a node. It is quantified by the ratio of the number of interconnections among the node's neighbors, and the number of maximum possible interconnections [20].

which is used to construct the topological time-series for links. We define time-series data for CN corresponding to link (x, y) in a time-window $w_{-\delta}$ as follows:

$$C_{-\delta}(x, y) = CN^{t-\Delta\delta}(x, y) - CN^{t-\Delta(\delta+1)}(x, y), \qquad (5)$$

where $CN^{t-\Delta\delta}(x, y)$ gives the common neighbor score between node-pair x and y respectively in $G^{t-\Delta\delta}$. We populate the values for $C_{-\delta}(x, y)$ in the sequence of contiguous time-windows that starts with the window where the first common neighbor of x and y appeared, and ends with w_0. This time-series is denoted as $C(x, y)$.

B. Modeling the time-series

The time-series defined in Subsection IV-A are used to forecast future trends. We apply *simple exponential smoothing* time-series forecasting method to model the time-series. The exponential smoothing method gives high importance to the recent activities, and importance decays exponentially from recent to less recent past. Let $Q_{-\delta}$ denotes the data corresponding to window $w_{-\delta}$ of a time-series Q. Forecast equation of the simple exponential smoothing model for Q can be given by following recurrence equation [19]:

$$Q'_{-\delta+1} = \alpha Q_{-\delta} + (1-\alpha) Q'_{-\delta}, \qquad (6)$$

where $Q'_{-\delta}$ gives the forecast value for time-window $w_{-\delta}$, given the time-series data present in previous time-windows; and $0 < \alpha \leq 1$ is called smoothing parameter. When $\alpha \to 0$, simple exponential smoothing gives same weightage to every window during forecast. As the value of α increases from 0 to 1, importance of the recent time-windows increases monotonically. As $\alpha \to 1$, importance of older windows decreases, and $\alpha = 1$ forecasts the value present in the last time-window. Equation 6 is used to forecast the value in the window w_1 representing the window just after time t, can be written as:

$$Q'_1 = \textstyle\sum_{i=0}^{q} \alpha(1-\alpha)^i \; Q_{-i} + (1-\alpha)^{q+1} Q'_{-q},$$

where $Q'_{-q} = Q_{-q}$ gives the first forecast, w_{-q} being the oldest window. α is estimated by minimizing the sum of the squared errors (SSE):

$$SSE = \textstyle\sum_{j=q-1}^{0} (Q_{-j} - Q'_{-j})^2 = \sum_{j=q-1}^{0} e_{-j}^2 = e^2,$$

where $e_{-j} = Q_{-j} - Q'_{-j}$ gives the error in window w_{-j}. Here, minimizing SSE is a nonlinear optimization task. We use vertical least square fitting procedure that estimates α by solving the equation:

$$\frac{d(e^2)}{d\alpha} = 0 \implies \textstyle\sum_{j=q-1}^{0} \dfrac{d(Q_{-j} - Q'_{-j})^2}{d\alpha} = 0.$$

C. Feature generation and unsupervised method

In this subsection, we propose an unsupervised method to predict dull nodes and links in a given network G^t, which are declared dull at a future time t'. Several features based on temporal change in dyadic and topological time-series are proposed here. These features capture the distinctive properties of a node (and a link) which is likely to be inactive or disappear, based on recent trends. As an example, from a topological perspective, if a node x does not make new connections lately, or there is a decreasing trend in making new connections, it may be a potential candidate of being dull in future. This is reflected in the time-series $d(x)$. Similarly, the trend in activity of a node x in recent times is reflected in the time-series $s(x)$. In case of links, if there is a decline in adding new common neighbors between the end nodes in recent time, they might break their relationship in near future. The forecast value of these time-series, generated using simple exponential smoothing model depicts how the nodes and links will behave in future in terms of the underlying node and link properties. We propose a number of features to predict dull nodes and links exploiting these forecast values. Table I summarizes the features proposed in this work. Suffix 1 in each of the feature values represents w_1, and the values are the forecast values for w_1 of the corresponding time-series (refer to Subsections IV-A and IV-B). $zeros_{node}$ and $zeros_{link}$ give the number of trailing zeros in dyadic time-series of the target node and the target link respectively. The $pref_{link}$ feature for a link is a combination of forecast values of the $degree$ time-series of the two end nodes. This feature is the temporal version of the preferential attachment property of real networks [21].

TABLE I. TIME SERIES DATA AND CORRESPONDING FEATURES.

	Value	Feature
Node x	$s'_1(x)$	$strength_{node}$
	—	$zeros_{node}$
	$d'_1(x)$	deg_{node}
	$c'_1(x)$	$clust_{node}$
Link (x, y)	$S'_1(x, y)$	$strength_{link}$
	—	$zeros_{link}$
	$d'_1(x)$	deg^{x}_{link}
	$d'_1(y)$	deg^{y}_{link}
	$d'_1(x) \times d'_1(y)$	$pref_{link}$
	$C'_1(x, y)$	$common_{link}$

A node or a link is described by its feature vector. *Min-Max normalization* of boundary $[0, 1]$ [3] is applied to all features. For simplicity, we use distance based unsupervised method to predict the dull nodes and links. We identify a *reference vector* that represents a dull node (and a dull link), and calculate its distance from each sample's feature vector. *Chebyshev distance* is used to measure the distance between the samples' feature vector and the reference vector. Chebyshev distance of the feature vector (say $\boldsymbol{X}$, where its elements are denoted as x_i's, i being integer values ranging from 1 to $|\boldsymbol{X}|$) and the reference vector (say $\boldsymbol{V}$, where its elements are denoted as v_i's) is calculated as follows:

$$D_{chebyshev}(\boldsymbol{X}, \boldsymbol{V}) = \max_{i=1}^{|\boldsymbol{X}|} |x_i - v_i|. \qquad (7)$$

$(1 - D_{chebyshev}(\boldsymbol{X}, \boldsymbol{V}))$ gives the sample's *proximity score* with a dull node (and a dull link). The values of all features other than $clust_{node}$, $zeros_{node}$ and $zeros_{link}$ of the reference vector are set to 0, because these features should have negative correlation with the dull ones. Corresponding

[3] It reassigns a feature value v of feature V as: $\frac{v - \text{minimum value in } V}{\text{maximum value in } V - \text{minimum value in } V}$

TABLE II. Dataset Specification

			γ	
Network	#Nodes	#Links	Node	Link
fb	44937	166511	13	18
dblp	1304128	5258985	14	10

values of $clust_{node}$, $zeros_{node}$ and $zeros_{link}$ are set to 1. After predicting the potential dull nodes and links, we remove them from G^t, for data cleaning task.

V. Datasets

Experiments are carried out on two time-stamped social networks of different characteristics: Facebook wall-post network [22] (fb), and DBLP co-authorship network (dblp). fb consists of all the wall-posts, posted in Facebook [4] New Orleans network spanning the period September 26th, 2006 and January 22nd, 2009. The data is available with Unix time-stamp representing the time of each wall-post, and the information of who is posting on whose wall. Each wall-post is considered as an interaction. We ignore the direction of interactions in order to prepare an undirected graph so that our method can be applicable on it. dblp dataset has been downloaded from the web-link http://dblp.uni-trier.de/xml/ in .xml format. It contains the publication information in the field of computer science from the year of 1936 upto the starting of 2014. Early years contain very few publications. So, in this paper we consider only the publications which occurs during the years $1980 - 2013$. Each publication information contains at least two kinds of tags $< author >$ and $< year >$, which represent each author and the year of the publication respectively. The type of a publication is characterized by tags like $< article >$ (for a journal publication), $< inproceedings >$ (for a conference proceedings publication), $< www >$ etc. Only the conference and journal publications are used to prepare the graph. We consider each publication as an interaction, which forms a clique among all authors of that publication, and a self-loop in case of single author paper. fb and dblp consider a *month* and a *year* long time-windows respectively. A summery of characteristics of the networks is presented in Table II.

VI. Dull nodes and links

Given a network G^t, after predicting the nodes and links which will be dull at a future time $t' = t + \Delta i$, we need to evaluate our method against true dull nodes and links in $G^{t'}$. The concept of dull nodes and links is subjective, which is defined as the nodes and links which has kept silent for long, probably will never interact in future. In this section, we propose a scheme to label the dull nodes and links by using the statistical property of the network $G^{t'}$.

To label a dull node, we consider the dyadic time-series ($s(x)$) of all nodes upto time-window w_0 (with respect to $G^{t'}$). A node is labeled dull if it has not interacted with others for a long time, *i.e.*, its dyadic time-series ends with large number of consecutive zeros. Subsequently, question arises; *how large that number should be, so that we label it as a dull node?* Here we define the number by exploiting the distribution of zero-burst (a series of consecutive zeros) patterns present in all nodes' time-series. It is explained using a toy example shown in Figure 1, which presents the time-series data of two nodes in $G^{t'}$. The first one from the top has all zeros in last eight windows; so it may be labeled dull in $G^{t'}$, if a dull node is defined by at least eight or more number of consecutive zeros at the end of the dyadic time-series. The second one is having two series of intermediate zeros (or zero-burst) of lengths one and five. The tail of the distribution of such zero-burst lengths over the time-series data of all nodes in $G^{t'}$ gives a view about *watching how many consecutive zeros, one can be confident enough that the node may not interact in future, ie., is dull.* Figure 2 shows the length distribution of intermediate zero-bursts for the dblp bibliographic network. Linearity shown in the plots, which are drawn in $log - log$ scale, indicates that the distributions follow power law. This finding supports the observation of [23], where the authors have found that *inter-event time* [5] distribution follows power law. Figure 2 indicates that, in real networks, the distribution function decreases rapidly with the increase of zero-burst length, and there are few cases of high number of intermediate zero bursts. This fact encourages us to label dull node as: let $F(z)$ be the *cumulative distribution function* of the intermediate zero-burst lengths in the time-series data, over all nodes for a given network $G^{t'}$. A node x in $G^{t'}$ is labeled dull iff its time-series data is having at least γ number of consecutive zeros at the end of its dyadic time-series such that, $F(\gamma - 1) \geq \beta$, where $0 < \beta < 1$ is a constant representing the confidence level. Without loss of generality, we set $\beta = 0.99$. Similarly, the class labels for *dull link* is labeled using the dyadic time-series data of all edges in $G^{t'}$. Table II presents γ values for dblp and fb.

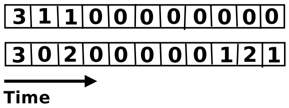
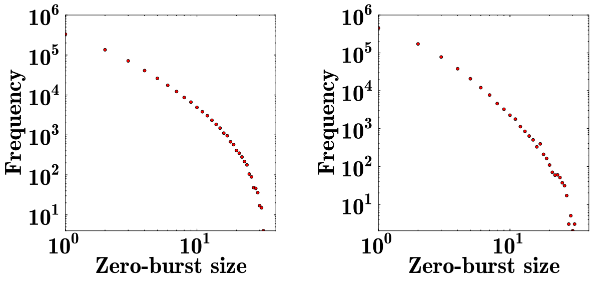

3 1 1 0 0 0 0 0 0 0 0
3 0 2 0 0 0 0 0 1 2 1

Fig. 1. A toy example.



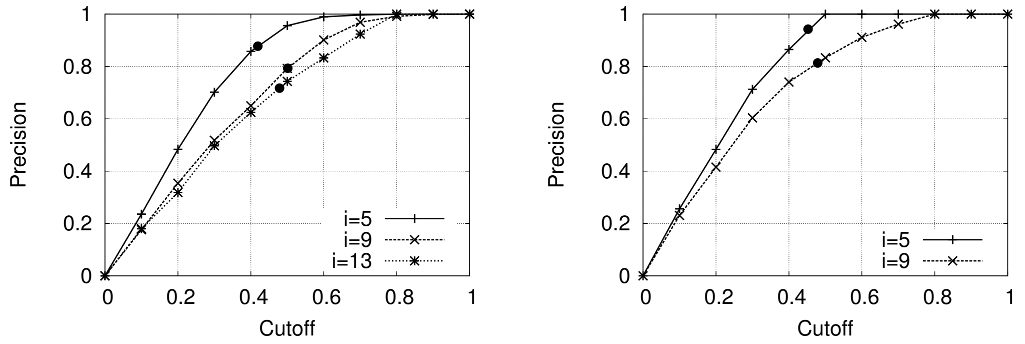
(a) Node (b) Link

Fig. 2. Log-log plot of the frequency of zero-burst patterns for DBLP bibliographic network.

Relation between dull nodes and links: By definition, dull nodes and dull links are closely related. Let, in a given network, γ_n represents the parameter γ to define dull nodes, and γ_l represents the parameter γ to define dull links. Let x be

[4] https://www.facebook.com/

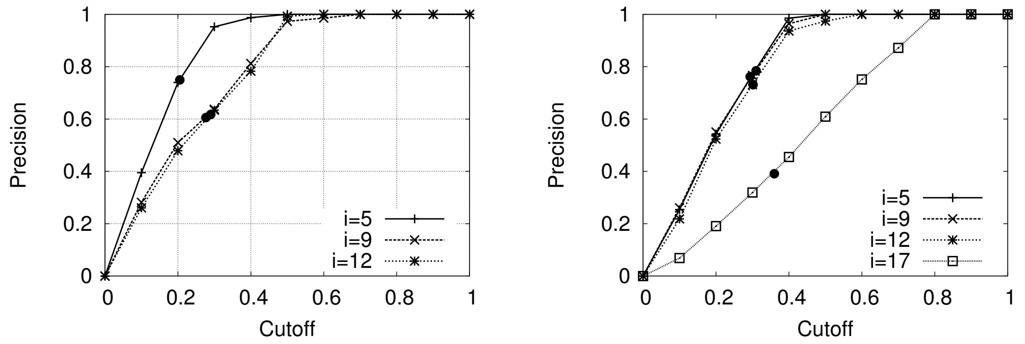
[5] Inter-event time is the time interval of occurrence of two consecutive events in a dynamic network.



(a) Node

(b) Link

Fig. 4. Precision curve for dull nodes and links prediction in dblp.



(a) Node

(b) Link

Fig. 5. Precision curve for dull nodes and links prediction in fb.

a node in the network. When $\gamma_n < \gamma_l$, if $s(x)$ has γ_n number of trailing zeros, it is possible that some of the links connecting x to its neighbors may not qualify as dull. Hence, predicting and pruning x removes those links too, which would not have been possible if predicting and pruning only the dull links was considered. Pruning the potential dull nodes adds an extra level of filtering.

VII. Evaluating proposed method

In this section, we evaluate the method proposed in Section IV. Experiments are performed over several snapshots: $G^{t'-\Delta i}, i < \gamma$, where $G^{t'-\Delta i}$ represents the given network G^t and the task is to predict the nodes and links which are declared dull at t'. Class imbalance is an inherent issue in this problem, because there are very few number of dull nodes and links as compared to the nodes and links which are not dull. We handle class imbalance by selecting only the nodes (and links) which have not interacted in window w_0 and w_{-1} (with respect to G^t) to represent the non-dull nodes (and links). In this paper, we pick top k number of samples in terms of their proximity score (described in Subsection IV-C) as potential dull nodes (and links), where k represents the number of true dull nodes (and links). Figure 3 presents the prediction performance in terms of *area under receiver operating characteristic (ROC) curve* (AUC scores) [24] for both datasets. High AUC scores in the results demonstrate that potential dull nodes and links can be predicted effectively. Prediction performance for the

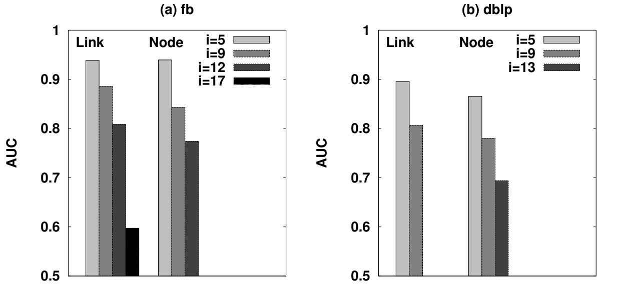


Fig. 3. Performance of dull nodes and links prediction at time $t' - \Delta i$ in terms of AUC score.

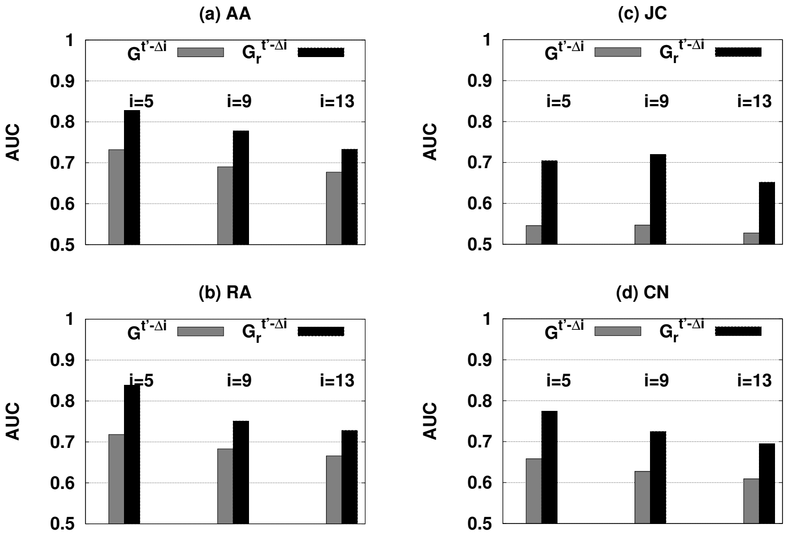


Fig. 6. Link prediction performance (dblp)

lower values of i is much better than the higher ones, because a high number of zeros at the end of all time-series of the dull samples boosts the performance when the value of i is small. Figures 4 and 5 present the prediction performance in different thresholds in terms of their *precision* scores. The black dots in each plot represent $precision@k$ values, k representing the number of true dull ones (the pruning threshold). The precision plots demonstrate that the proposed method achieves very high precision before reaching 0.5 cutoff value. As we set k as the cutoff in our prediction mechanism, $precision@k$ will govern the performance of our method in practical scenario, which is basically a retrieval process. The plots show that our method can achieve high $precision@k$ values. Although rise of the precision curves for fb is more rapid than dblp, $precision@k$ values win in dblp. It is also notable that, in spite of overcoming class imbalance, the proportion of dull nodes and links are much less in fb, which is another cause of performance degradation at the retrieval point.

VIII. Data cleaning and link prediction: a case study

After predicting the dull nodes and links from $G^{t'-\Delta i}$, we investigate the effect of cleaning the datasets by removing those nodes and links, on identifying new links that will appear in $G^{t'}$. We first sort the proximity scores for all samples (links and nodes) and consider a number of best performing k samples, where k represents the number of labeled dull nodes and links, as the potential candidate of being dull. We prepare a graph $G_r^{t'-\Delta i}$ by removing the predicted dull nodes and links from $G^{t'-\Delta i}$. We perform link prediction on both of $G^{t'-\Delta i}$ and $G_r^{t'-\Delta i}$ using state-of-the-art link prediction methods, and compare their performance in terms of AUC scores. We consider four baseline link prediction methods: *Common neighbor (CN), Jaccard's coefficient (JC), Adamic/Adar (AA), and Resource allocation (RA)* [7], [25]. These methods are described in Table III. We perform the experiments for multiple values of i. Figures 6 and 7 summarizes the link prediction results for dblp and fb. These figures show that prediction performance in the reduced graph $G_r^{t'-\Delta i}$ improves significantly for almost all values of i's over both the datasets. dblp network is benefited more than fb by data cleaning. It can be justified by its higher $precision@k$ values in dull link prediction (refer to Section VII). In most of the cases, over all link prediction methods and datasets, effect of data cleaning increases with increase in the value of i. In fb, sometimes data cleaning worsens the link prediction when $i = 17$. It is expected because, as shown in Figure 5, $precision@k$ value for dull link prediction is as low as .39 in $G^{t'-17\Delta}$. Among the link prediction methods, JC and CN are the most affected prediction methods, whereas AA and RA are more robust against noise.

TABLE III. State-of-the-art link prediction methods.

Method	Definition
CN	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
JC	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
AA	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log\lvert\Gamma(z)\rvert}$
RA	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\lvert\Gamma(z)\rvert}$

Note: $\Gamma(a)$ denotes the set of neighbors of node a.

IX. Conclusion

In this paper, we have proposed a time aware network data cleaning method. The proposed method predicts dull nodes and links, based on historical network properties. It then prunes them from the network. This paper has also proposed a scheme to label dull nodes and links for evaluating the proposed method. Exhaustive experiments on two real network datasets have shown that the proposed method can effectively predict the potential dull nodes and links. It further

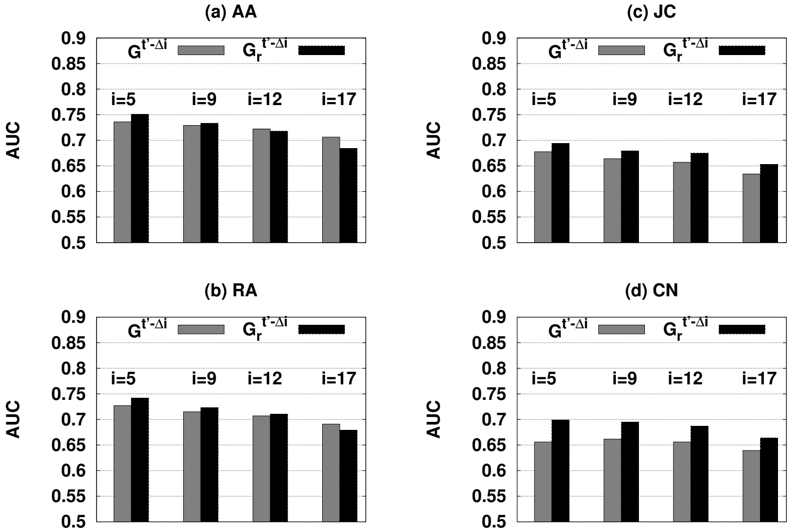


Fig. 7. Link prediction performance (`fb`)

has demonstrated the effect of network data cleaning on state-of-the-art link prediction methods. Significant improvement in link prediction performance has been observed when the proposed data cleaning method is applied on real datasets.

REFERENCES

[1] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications*, vol. 311, no. 3, pp. 590–614, 2002.

[2] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[3] T. Tylenda, R. Angelova, and S. Bedathur, "Towards time-aware link prediction in evolving social networks," in *SNA-KDD '09*. New York, NY, USA: ACM, 2009, pp. 9:1–9:10.

[4] E. Richard, S. Gaïffas, and N. Vayatis, "Link prediction in graphs with autoregressive features," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 565–593, Jan. 2014.

[5] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 554–560.

[6] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng, "On evolutionary spectral clustering," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 4, pp. 17:1–17:30, dec 2009.

[7] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," in *Conference on Information and Knowledge Management (CIKM03)*, 2003, pp. 556–559.

[8] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[9] J. Zhang, Y. Wang, and J. Vassileva, "Socconnect: A personalized social network aggregator and recommender," *Information Processing & Management*, vol. 49, no. 3, pp. 721–737, 2013.

[10] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, 2010, p. 12.

[11] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava, "Prediction promotes privacy in dynamic social networks," in *Proceedings of the 3rd conference on Online social networks*, 2010, pp. 6–6.

[12] A. A. Ferreira, M. A. Gonçalves, and A. H. Laender, "A brief survey of automatic methods for author name disambiguation," *Acm Sigmod Record*, vol. 41, no. 2, pp. 15–26, 2012.

[13] M. Huisman and C. Steglich, "Treatment of non-response in longitudinal network studies," *Social networks*, vol. 30, no. 4, pp. 297–308, 2008.

[14] C. Hernández and G. Navarro, "Compressed representations for web and social graphs," *Knowledge and information systems*, vol. 40, no. 2, pp. 279–313, 2014.

[15] S. A. Macskassy, "Mining dynamic networks: The importance of pre-processing on downstream analytics," *COMMPER 2012*, p. 2, 2011.

[16] K. Y. Kamath and J. Caverlee, "Transient crowd discovery on the real-time social web," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 585–594.

[17] T. Raeder, O. Lizardo, D. Hachen, and N. V. Chawla, "Predictors of short-term decay of cell phone contacts in a large scale communication network," *Social Networks*, vol. 33, no. 4, pp. 245–257, 2011.

[18] G. Miritello, *Temporal patterns of communication in social networks*. Springer Science & Business Media, 2013.

[19] R. Hyndman, A. Koehler, J. Ord, and R. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*, ser. Springer Series in Statistics. Springer, 2008.

[20] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[21] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical review E*, vol. 64, no. 2, p. 025102, 2001.

[22] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *WOSN '09*. New York, NY, USA: ACM, 2009, pp. 37–42.

[23] M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész, "Universal features of correlated bursty behaviour," *Scientific reports*, vol. 2, 2012.

[24] J. A. Hanley and B. J. Mcneil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, april 1982. [Online]. Available: http://radiology.rsnajnls.org/content/143/1/29.abstract

[25] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 71, no. 4, pp. 623–630, 2009.