**Paper No: SC-04** **Smart Computing**

# Temporal preferential attachment: Predicting new links in temporal social networks

Panchani Wickramarachchi*
*Software Engineering Teaching Unit*
*Faculty of Science, University of Kelaniya, Sri Lanka*
panchaniwickramarachchi@gmail.com

Lankeshwara Munasinghe
*Software Engineering Teaching Unit*
*Faculty of Science, University of Kelaniya, Sri Lanka*
lankesh@kln.ac.lk

***Abstract* - Social networks have shown an exponential growth in the recent past. It has estimated that nearly 4 billion people are currently using social networks. The growth of social networks can be explained using different models. *Preferential Attachment (PA)* is a widely used model, which is often used to link prediction in social networks. PA tells that the social network users prefer to get linked with popular users in the network. However, the popularity of a node depends not only on the node's degree but also on the node's activeness which is reflected by the amount of active links the node has at present. Activeness of a link can be quantified using the timestamp of the link. The present work introduces a novel method called *Temporal Preferential Attachment (TPA)* which is defined on the activeness and strength of a node. Strength of a node is the sum of weights of links attached to the node. Here, the weights of the links are assigned according to their activeness. Thus, TPA captures the temporal behaviors of nodes, which is a vital factor for new link formation. The novel method uses *min - max* scaling to scale the time differences between current time and the timestamps of the links. Here, the *min* value is the earliest timestamp of the links in the given network and *max* value is the latest timestamp of the links. The scaled time difference of a link is considered as the *temporal weight* of the link, which reflects its activeness. TPA was evaluated in terms of its link prediction performance using well-known social network data sets. The results show that TPA performs well in link prediction compared to PA, and show a significant improvement in prediction accuracy.**

***Keywords - activeness of links, link pre- diction, social networks, TPA***

## I. INTRODUCTION

At present, around 4 billion users are using social networks, and still the number grows exponentially. Social networks serve different interests of the users. For example, social networks such as Facebook serve mainly as a friendship network which allow users to share their content and thoughts with their friends. In contrast, question and answering social networks such as Stackoverflow serve users to solve their programming problems by sharing them with other users of the social network. In addition, opinion posting social networks such as Reddit and Slashdot provide users a platform to post their opinions, thoughts, views and comments on various topics. Therefore, the growth of each social network depends on different facts and hence, predicting the growth of social networks has become a complex task [1], [2]. A plethora of researches have been carried out to devise novel models or alter the existing models to describe the growth of complex and heterogeneous social networks.

Social networks present a picture which has users connected via links. This picture of social networks can further elaborate as a set of nodes connected via single or multiple edges (In network theory terminology, the users are referred to as nodes and the links referred to as edges). Here, the multiple edges represent the interactions that happen between the node pairs. For example, in Facebook, once a pair of users become friends, they interact with each other in multiple ways such as chatting, commenting, sharing posts, etc. All these interactions are considered as temporal edges and hence, the words edge and interaction use interchangeably to refer to the same entity. In network theory, the number of interactions between a node pair is referred to as the edge weight which reflects the closeness of the node pair. The total of the weights of edges attached to a node is said to be the strength of the node. In other words, the degree of the node is considered as the strength of a node. Here, the node degree is the count of all temporal edges attached to the node. The strength of a node reflects its popularity in the social network. The higher the strength, the higher the popularity. However, this is not always true due to the temporal behavior of nodes and edges. In other words, the strength of a node varies over time due to various factors. Therefore, the present research investigates the primary causes of temporal behavior of social networks. Although this study focuses on online social networks, it can be generalized to other types of social networks as well. The contribution of this paper can be summarized as follows.

- Provide an insight about the temporality of social networks.
- Discuss the limitations of existing static features used for link prediction in social networks.
- Introduce a non-parametric time-aware feature, *Temporal Preferential Attachment (TPA)* which captures the temporal behavior of nodes and edges.

The rest of the paper is organized as follows. Section II discusses the related research and provides a better insight about the importance of studying the temporality of social networks for link prediction. Section III presents the details of TPA, and link prediction performance of TPA. Section IV contains the experimental evaluation of the new method. Finally, section V concludes the paper with the summary of the research and future directions.

## II. RELATED RESEARCH

Modeling modern social networks is a formidable task due to their complexity, heterogeneity and the size. Past researches have introduced various models to describe the growth of social networks [3], [4]. A growth model is a set of rules or a theory by which new nodes and edges are added to a social network. Among those growth models, the *Preferential Attachment (PA)* is a widely used method,

which is often used for link prediction in social networks. The intuition behind PA is that the nodes of social networks prefer to get linked with higher degree nodes or the popular nodes. PA quantifies this preference on popular nodes. Out of various PA based growth models, this section reviews some of the popular PA based growth models.

Baraba´si-Albert (BA) model [5] tells that the social networks grow according to the so-called power law (see Equation 7). The network starts with $n$ nodes connected each other and grows by adding new nodes where each new node $v$ randomly finds an existing node $u$ to connect according to the probability proportional to the degree of $u$ (see Equation 1).

$$\prod(d_u|v) = \frac{d_u}{\sum_{i \in N} d_i} \quad (1)$$

where $N$ is the set of nodes in the network and $d_u$ is the degree of node $u$. Although the BA model works well in modeling technological networks such as the Internet, it shows some limitations in modeling modern social networks such as friendship networks. The probability or the preference of choosing a node to connect does not depend only one the degree distribution of the nodes in the network but there are some other factors such as homophily, node attributes, and node activeness. Among them, homophily is described as the preference of new nodes to get linked with nodes which have similar interests. Considering this characteristic, homophily model [6] was introduced with homophily parameter $\delta$ which quantifies a certain property of a node. For any node pair $u$ and $v$, the homophily parameters are defined as $u_\delta$ and $v_\delta$. The difference $\Delta_{uv} = |u_\delta - v_\delta|$ tells the closeness of the node pair. Thus, the connection probability is defined as:

$$\prod(d_u|v) = \frac{(1-\Delta_{uv})\, d_u}{\sum_{i \in N} (1-\Delta_{iv})\, d_i} \quad (2)$$

Homophily model improves BA model by incorporating the similarity between node properties. Thus, the homophily model shows better performance in modelling modern social networks such as friendship networks. However, it still falls short in capturing temporality of nodes which is a key factor in deciding the connection probability. Therefore, an alternative model called Fitness model [7] was introduced to capture the short term node popularity. Fitness model is similar to BA model, but it includes an additional parameter called fitness parameter $\eta$ $(0 \leq \eta \leq 1)$ which captures the short term popularity of the node. The connection probability of Fitness model is defined as:

$$\prod(d_u|v) = \frac{\eta_u d_u}{\sum_{i \in N} \eta_i d_i} \quad (3)$$

Although the Fitness model captures the node temporality, it is still required to estimate the fitness parameter for each network. As a consequence, this model cannot generalise across different social networks. Also, parameter estimation is computationally intensive. Due to those limitations, researchers have introduced non-parametric link prediction methods. Non-parametric link prediction algorithm (NonParam) [8] uses a sequence of graph snapshots over time to capture the dynamic behavior of nodes and edges. Compared to the baselines (Last time of linkage, Common neighbors, Adamic/Adar and Katz), NonParam algorithm performed well even in the presence of seasonal patterns. However, it can only predict pairs which are generated by 2-hop neighborhoods of last timesteps. Moreover, the non-parametric latent feature relational model is another link prediction method used to infer the latent binary features in relational entities [9]. This method has used feature-based methods to analyze the network data with the idea of Bayesian non-parametric approach. In capturing the subtle patterns of interactions, the latent relational model has performed better than class-based models.

Apart from that, researchers have introduced growth models which consider structural patterns such as motifs in temporal social networks [10], patterns and dynamics of users' behavior and interaction in social networks [11]. Inclusion of location information into PA based models have shown significant improvement in modeling the growth of various social networks [12]. This research has introduced a growth model which captures the growth of population in different geographic locations. It considers the account creation time and geographic information of each user. Although the above approaches have shown promising results in modeling the growth of modern social networks, still they have their own limitations.

## III. Link Prediction in Social Networks

Link prediction in social networks is a well-established research area. Social networks grow by adding new nodes as well as new links. Therefore, knowing the growth pattern of a social network is essential for link prediction in social networks. Link prediction problems can be classified into several sub-problems. For example, predicting new links, predicting missing links and hidden links are the popular link prediction tasks. This research focused on new link prediction, which can be defined as follows. For a given network at time $t$ our task is to predict the potential links that can appear in time $t + 1$ [13]. Emergence of new links depends on various factors such as structural features, similarities between node and edge attributes. Common neighbors, Jaccard's coefficient, Adamic/Adar index, and PA are a set of popular neighbors based structural features used for link prediction [14]. Among them, PA quantifies this preference of getting linked with popular nodes. For example, preference of node pair $ii$ and $j$ getting linked can be quantified as shown in Equation 4.

$$PA_{ij} = degree_i \times degree_j \quad (4)$$

where $degree_i$ is the degree of node $i$. For example, in Figure 1, node *A* has degree 4 and node *B* has degree 3. Therefore, the $PA_{AB} = 12$. According to Equation 4, if the nodes have higher degrees their PA score takes a higher value. In case of link prediction, node pairs with higher PA are highly likely to get linked in future. Although PA looks like a promising method for link prediction based on the node popularity, the limitation of PA is it assumes that the popularity of a node solely depends on the node degree. In other words, the strength of the node, which assigns an equal weight (one) for each edge irrespective of its activeness. However, the popularity of a node depends not only on the node's degree but also on the activeness of the node which is reflected by the amount of active edges the

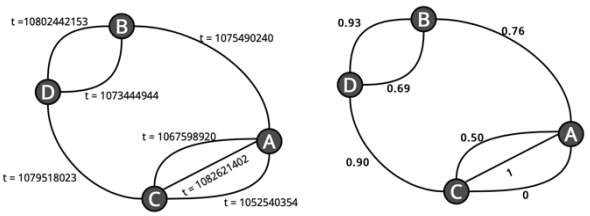
node has at present. In other words, the activeness of the node is reflected by the amount of recent interactions with its neighbors. The activeness of those edges is relatively higher than the old edges (old interactions). Thus, Activeness of an edge can be quantified using the timestamp of the edge. Based on the edge activeness, some of the recent researches have introduced alternative time-aware features which have shown their success in link prediction in social networks [15]– [17]. However, the inherent problems of most of these time-aware features are that they include parameters. Thus, it is required to optimize the parameters to obtain the optimal results. Parameter optimization is a tedious task as it consumes time and large amounts of computational power. As a consequence, some of those time-aware methods cannot generalise across different social networks. Those limitations motivated us to introduce a novel non-parametric time-aware feature which is an alternative to PA.

### *A. Temporal Preferential Attachment*

The present work introduces a novel method called *Temporal Preferential Attachment (TPA)* which is defined on the strength or the weighted node degree where the weights of the edges are assigned according to the activeness of the links. Thus, TPA captures the temporal behaviors of nodes, which is a vital factor for new link formation. The novel method uses $min - max$ $min - max$scaling to scale the time differences between current time and the timestamps of the links. Here, the $minmin$ value is the earliest timestamp of the links in the given network and $max$ value is the latest timestamp of the links. The scaled time difference of an edge is considered as the *temporal weights* (see Equation 5) of the link, which reflects its activeness.

$$Temporal\ weight_{ij} = \frac{T_{ij} - T_{min}}{T_{max} - T_{min}} \quad (5)$$

where $T_{ij}$ is the timestamp of the edge $ij$, $T_{max}$ is the latest timestamp in the network and $T_{min}$ is the earliest timestamp.




(a) (b)

Fig. 1. A temporal social network. Figure (a): edges assigned with timestamps. Figure (b): after scaling the timestamps, each edge is assigned with a temporal weight.

According to Figure 1, older edges get lower weight and recent edges get higher weight. This is far better than assigning equal weights to all edges because the temporal weights reflects the activeness of the edges and hence, the activeness of the nodes they attached. Based on the temporal weights, TPA of nodes $i$ and $j$ calculate as shown in Equation 6.

$$TPA_{ij} = TS_i \times TS_j \quad (6)$$

where $TS_i$ is the *temporal strength* of node $i$. Temporal strength of a node is defined as the total of temporal weights of the edges attached to the node. In Figure 1b, temporal strength of node *A* is 2.26 and temporal strength of node *B* is 2.38. Therefore, $TPA_{AB}$= 5.38 which is less than $PA_{AB}$ but better captures the temporal strengths of the node pair. The effectiveness of novel method TPA was tested in terms of its link prediction performances on real-world social networks.

## IV. Experimental Analysis

The present study specifically focuses on link prediction in question and answering social networks and opinion posting social networks. In addition, one online friendship network was also used in the experiments to compare the effectiveness of TPA against PA in different settings. There are three types of interactions in question and answering networks: answers to the questions, comments to the questions, and comments to the answers. In this experimental analysis, we disregard the type of the interaction and consider each interaction as a temporal edge. TPA was evaluated in terms of its link predicting performances. The performance metric used to compare PA and TPA was area under curve (AUC) and ROC curves which give a better picture in model comparison.

The data analytics show that their degree distributions of the six networks follow the notion of power law (see Figure 2) which says that the fraction $P(k)$ of nodes in the network having degree $k$ goes for large values of $k$ according to the Equation 7.

$$P(k) = \lambda k^{-\gamma} \quad (7)$$

Here, γ is a parameter which typically takes values in between 2 and 3 for scale-free networks.

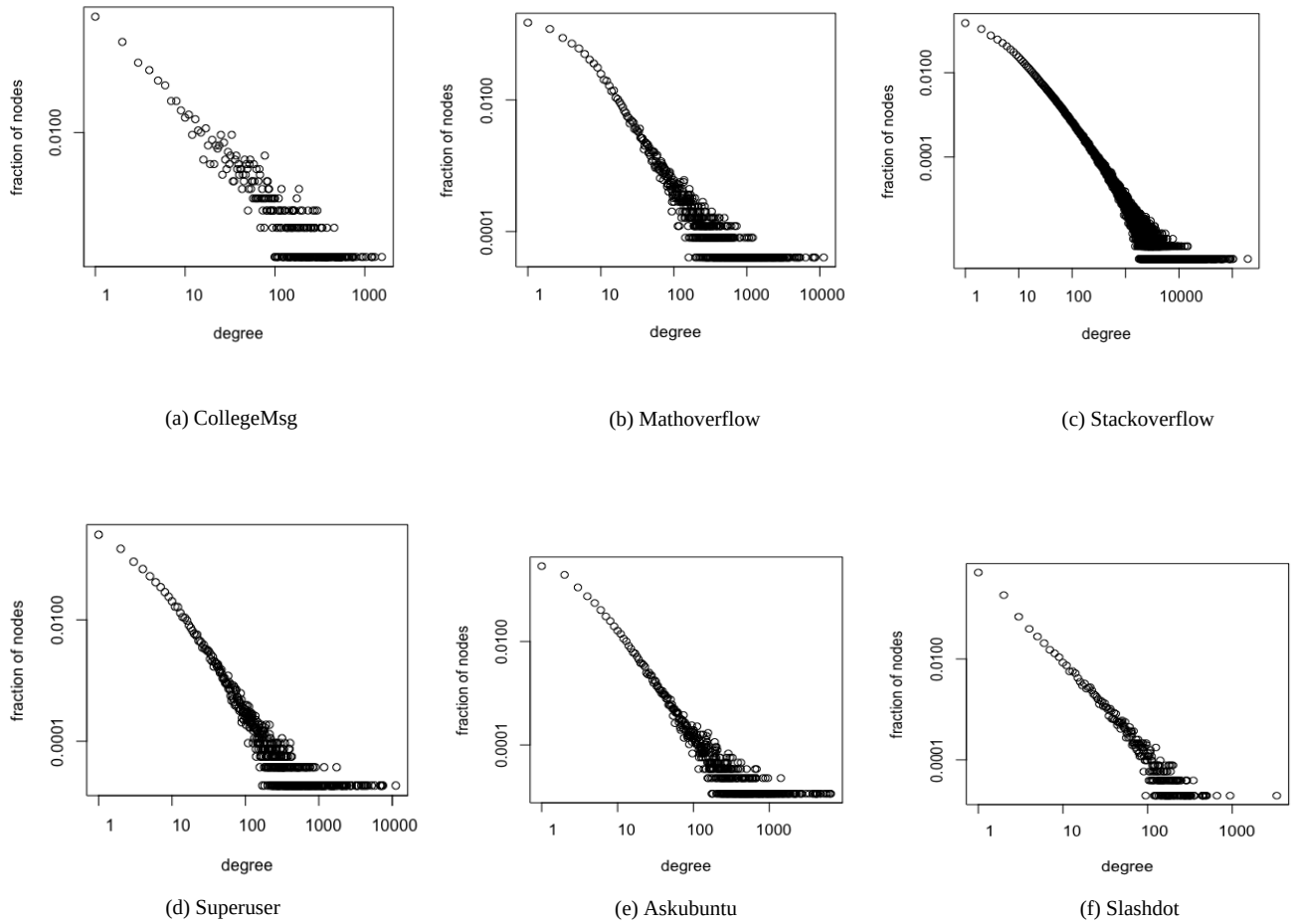
### *A. Data*

Four question and answering social network data sets, one opinion posting social network data set and one online social network data set were used to test the effectiveness of TPA. Summary statistics of the data sets are shown in Table I. All data sets used in the experiment were taken from Stanford Large Network DataSet Collection (https://snap.stanford.edu/data/).

To create training sets and test sets, each data set was sorted in the ascending order of timestamps, and 80% of the sorted data set was taken as the training set and the rest 20% with latest timestamps were taken as the test set. In addition, all networks were assumed undirected. In each network, the largest connected subgraph was used to test the link prediction performance of PA and TPA. The training and test graphs were created in a way that the positive examples are the edges which are present in the test graph but not present in the training graph, and the negative examples are the non-edges which are common to training and test graphs. Also, all the nodes in the test graph are present in the training graph.

TABLE I. STATISTICS OF THE NETWORKS

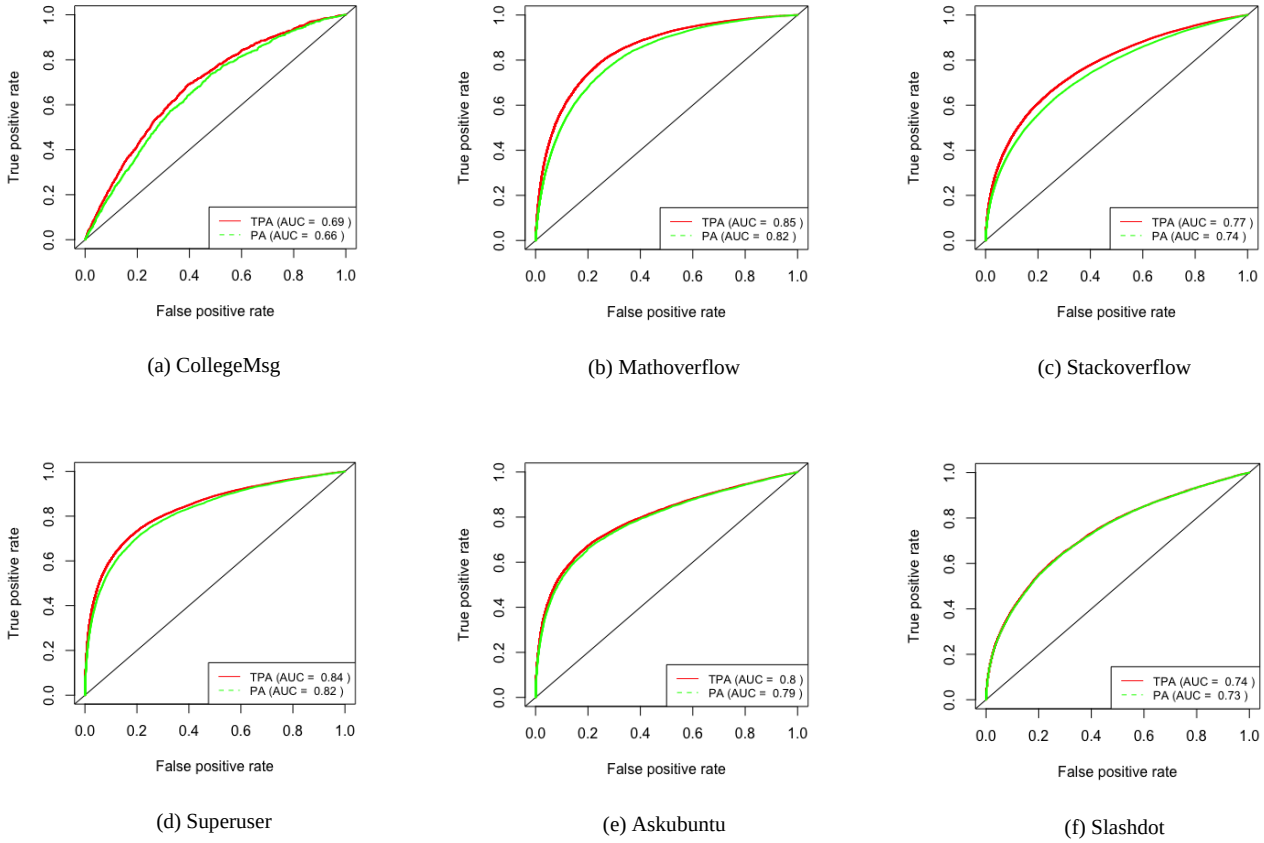
| **Network feature** | **CollegeMsg** | **Mathoverflow** | **Stackoverflow** | **Superuser** | **Askubuntu** | **Slashdot** |
|---|---|---|---|---|---|---|
| Nodes | 1899 | 24818 | 23977 | 53657 | 87485 | 51083 |
| Edges | 59835 | 506550 | 500000 | 500000 | 500000 | 140778 |
| Time Span (days) | 194 | 2305 | 201 | 1350 | 1875 | 13395 |
| Nodes in Largest WCC | 1893 | 24668 | 23906 | 52477 | 83497 | 51083 |
| Edges in Largest WCC | 59831 | 506395 | 499920 | 498942 | 496603 | 140778 |
| Average clustering coefficient | 0.11 | 0.31 | 0.08 | 0.12 | 0.1 | 0.02 |
| Number of triangles | 14319 | 1403919 | 849247 | 704332 | 371319 | 18937 |
| Diameter (Longest shortest path) | 8 | 10 | 10 | 13 | 13 | 17 |
| Density | 0.03 | 0.00164 | 0.00174 | 0.00035 | 0.00013 | 0.00011 |




(a) CollegeMsg (b) Mathoverflow (c) Stackoverflow

(d) Superuser (e) Askubuntu (f) Slashdot

Fig. 2. Degree Distribution




(a) CollegeMsg (b) Mathoverflow (c) Stackoverflow

(d) Superuser (e) Askubuntu (f) Slashdot

Fig. 3. Model comparison: ROC Curves of PA and TPA

TABLE II: LINK PREDICTION PERFORMANCE OF PA AND TPA. AUC COMPARISON OF PA AND TPA.

| Network | AUC of TPA | AUC of PA |
|---|---|---|
| CollegeMsg | **0.69** | 0.66 |
| Mathoverflow | **0.85** | 0.82 |
| Stackoverflow | **0.77** | 0.74 |
| Superuser | **0.84** | 0.82 |
| Askubuntu | **0.80** | 0.79 |
| Slashdot | **0.74** | 0.73 |

### *B. Results*

The summary of the results of the experimental analysis is shown in Table II. It shows that TPA performs better than PA in link prediction in all six social networks. Among them, TPA shows 3% improvement in link prediction accuracy on Mathoverflow, Stackoverflow and CollegeMsg networks. TPA reports 2% improvement in link prediction accuracy on Superuser network. In Askubuntu and Slashdot networks, TPA reports 1% improvement in link prediction accuracy over PA. These results revealed that TPA performs well on most of the question and answering networks. The activeness of the nodes in question and answering networks stays for a short period of time. Once the question gets the right answer, all the interactions with that node stops, and the node becomes inactive. Then the new links start to emerge around new questions rather than older ones. Owing to this nature, TPA performs better than PA in link prediction.

## V. DISCUSSION AND CONCLUSION

Modelling the growth of social networks is a challenging task due to various factors. Among them, the temporality of nodes and edges is a key factor which influences the emergence of new edges. This research introduced a simple yet effective growth model TPA based on the node activeness. The underneath assumption of TPA is each node $vv$ randomly finds an existing node $u$ $u$to connect according to the probability proportional to the temporal strength of $u$ (see Equation 8).

$$\prod(TS_u|v) = \frac{TS_u}{\sum_{i \in N} TS_i} \tag{8}$$

Here, $TS_u$ is the temporal strength of node $u$. This growth model somewhat similar to the Fitness model [7]. The key difference is that the Fitness model includes a parameter but the TPA based growth model is non-parametric model. This growth model can be further improved by incorporating homophily and node attributes, which is the future direction of this research.

Although the novel growth model assumed that social networks obey the scale-free property, most of these real world networks do not follow the power law (see Equation 7). Among the social networks used in this study, degree

distributions of Superuser and Askubuntu follow the power law with the exponent of $\gamma = 2.1$. However, degree distributions of Mathoverflow and Slashdot follow the power law with the exponents less than two ($\gamma = 1.7$ and $\gamma = 1.9$). In Stackoverflow and CollegeMsg networks the power law exponents are 3.1 and 3.9 respectively. Typically, the γ of scale-free networks lies in between 2 and 3. The γ value of four above real-world networks stay outside the typical range, which mean that those networks are not typical scale-free networks. According to the Figure 2, the fraction of higher node degrees ($1000 \leq$ degree) are much higher in Askubuntu, Math overflow, Stackoverflow and Superuser networks. It reflects the fact that those networks are growing around the higher degree nodes. Thus, the growth mechanisms of those networks might not fully explained by the power low assumption but still TPA growth model performs better than PA growth model.

Activeness of a node reflects by its interactions with its neighbors. Frequent and recent interactions make the node active. If the node is active then it should make two-way interactions with its neighbors. Otherwise, if the interactions are one-way, which means neighbors to node then the activeness of the node is questionable. In other words, the neighbors interact with the node but the node is not interacting with any of its neighbors. In this case, the node cannot be regarded as an active node. The present research considered both one-way and two-way interactions make the node active. However, it is required to investigate the one-way interactions and two- way interactions separately because in the one-way case only the edge is active but the node might not active. Therefore, it requires thorough investigation about different types of interactions to understand the insights of activeness.

Although TPA shows its own limitations, it shows better performance in link prediction compared to PA. Specially, TPA shows impressive performance over the temporal social networks. In fact, TPA is an effective non-parametric model which can be used to model the temporal social networks as well for link prediction.

## REFERENCES

[1] O. Mokryn, A. Wagner, M. Blattner, E. Ruppin, and Y. Shavitt, "The role of temporal trends in growing networks," *PLOS ONE*, vol. 11, p. e0156505, 08 2016. [Online]. Available: http://www.cs.tau.ac.il/ ruppin/temporal.pdf

[2] F. Gao, K. Musial, C. Cooper, and S. Tsoka, "Link prediction methods and their accuracy for different social networks and network metrics," Scientific Programming, 2015.

[3] G. G. Piva, F. L. Ribeiro, and A. S. Mata, "Networks with growth and preferential attachment: Modeling and applications," 2020.

[4] M. Newman, "Newman, m.e.j.: Clustering and preferential attachment in growing networks. phys. rev. e 64, 025102," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 64, p. 025102, 09 2001.

[5] A.-L. Barabasi and R. Albert, "Albert, r.: Emergence of scaling in random networks. science 286, 509-512," *Science (New York, N.Y.)*, vol. 286, pp. 509–12, 11 1999.

[6] M. Almeida, G. Mendes, G. Madras, and L. Silva, "Scale-free homophilic network," *The European Physical Journal B*, vol. 86, 02 2013.

[7] G. Bianconi and A.-L. Barabasi, "Competition and multiscaling in evolving networks," *EPL (Europhysics Letters)*, vol. 54, p. 436, 05 2001.

[8] P. Sarkar, D. Chakrabarti, and M. Jordan, "Nonparametric link prediction in large scale dynamic networks," *Electronic Journal of Statistics*, vol. 8,01 2014.

[9] K. Miller, M. Jordan, and T. Griffiths, "Nonparametric latent feature models for link prediction," in *Advances in Neural Information Pro- cessing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., vol. 22. Curran Associates, Inc., 2009.

[10] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, Feb 2017. [Online]. Available: http://dx.doi.org/10.1145/3018661.3018731

[11] P. Panzarasa, T.Opsahl, and K.Carley, "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community," JASIST, vol. 60, pp. 911–932, 05 2009.

[12] K. Zhu, W. Li, and X. Fu, "Modeling population growth in online social networks," *Complex Adaptive Systems Modeling*, vol. 1, 12 2013.

[13] D. Liben-nowell and J. Kleinberg, "The link prediction problem for social networks," Journal of the American Society for Information Science and Technology, vol. 58, 01 2003.

[14] L. Lu¨ and T. Zhou, "Link prediction in complex networks: A survey," Physica A: Statistical Mechanics and its Applications, vol. 390, no. 6, pp. 1150–1170, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037843711000991X

[15] L. Munasinghe and R. Ichise, "Time score: A new feature for link prediction in social networks," *IEICE Transactions on Information and Systems*, vol. E95.D, p. 821–828, 2012.

[16] L. Munasinghe and R. Ichise, "Link prediction in social networks using information flow via active links," *IEICE Transactions on Information and Systems*, vol. E96.D, pp. 1495–1502, 2013.

[17] E. Bu¨tu¨n, M. Kaya, and R. Alhajj, "Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks," *Information Sciences*, vol. 463-464, pp. 152–165, 2018.