Hybrid Structure–based Link Prediction Model

Fei Gao
King's College London, UK
Email: fei.1.gao@kcl.ac.uk

Katarzyna Musial–Gabrys
Bournemouth University, UK
Email: kmusialgabrys@bournemouth.ac.uk

***Abstract*—In network science several topology–based link prediction methods have been developed so far. The classic social network link prediction approach takes as an input a snapshot of a whole network. However, with human activities behind it, this social network keeps changing. In this paper, we consider link prediction problem as a time–series problem and propose a hybrid link prediction model that combines eight structure-based prediction methods and self-adapts the weights assigned to each included method. To test the model, we perform experiments on two real world networks with both sliding and growing window scenarios. The results show that our model outperforms other structure–based methods when both precision and recall of the prediction results are considered.**

I. Introduction

The rapid development of the Internet has pushed the research in the area of network science to the entirely new level. More and more human activities have been moved from off-line to on-line world and this resulted in vast amount of data available for investigation. Online social networks, ranging from collaboration networks to friendship networks have been widely studied by researchers from different fields. These social networks can be represented as graphs where nodes are users and links indicate social interactions between those users. Driven by human activities, social network keeps changing which makes the network prediction a challenging and worth studying topic.

Our work focuses on the link prediction problem formalized in [1]. The classic approach for solving the link prediction problem is first to take a snapshot of a network resulting from the time frame $[t_0, t_1]$. New links are predicted based on the network topology existing in $[t_0, t_1]$. The results are verified with the real world network snapshot from the period $(t_1, t_2]$. Algorithms for links prediction typically compute similarity score between two nodes and assume that nodes with larger scores are more likely to be connected in the future. Link prediction problem has been studied on various networks such as disease spread networks [2], [3], [4], scientific collaboration networks [5], [6] and online social networks [7], [8], [9]. Existing research has shown that some prediction methods perform well on networks with specific characteristics. For instance, in [5], authors found that the Katz and Preferential Attachment methods work well on a book sales recommendation network. Authors in [6] claimed that Adamic/Adar method provides the best prediction accuracy on Wikipedia Collaboration Network. The issue is that the performance of methods relies much on the network topology [10]. A prediction method that could self-adapt to different networks is thus required. In addition, the traditional link prediction study approach takes network as a static graph by using network snapshot which cannot reflect the continuous evolution of social networks dynamics. In this paper we introduce a hybrid link prediction model. Data used in this model are time-stamped. We apply two approaches: (i) sliding and (ii) growing window when splitting the data for analysis. The proposed hybrid model combines eight widely used topology based link prediction methods with the assumption that networks evolve following certain mechanisms (we call them rules). Our model predicts links based on the rules that we learn from the past data about the network. The model has been tested with two real world social networks, Facebook friendship network and Wroclaw University of Technology email communication network. The results show that the hybrid model performs better than the other eight methods applied separately. It is also shown that the two analysed networks are evolving in different ways.

The rest of the paper is organised as follows: in Section II, we introduce the hybrid model as well as we present methods that were combined in the hybrid model. Section III describes the design of the experiments. Following this, we discuss the results of the experiments in Section IV. The last Section 5 concludes the findings and ideas for future work are presented.

II. Hybrid Link Prediction Model

Much effort has been made to develop new link prediction methods and many of those methods have been proved to perform well on different networks topologies. There is no prediction method that performs well for all networks [10]. Many of the existing prediction methods work better if the network is growing following the same mechanism over time. For example, the common neighbour approach assumes that links are more likely to appear between nodes with more common neighbours. Only if the network evolves following this rule the common neighbours prediction model will give better prediction accuracy than other methods. This applies to other prediction methods as well, e.g. preferential attachment approach. However, a real world network might not evolve following only one rule; it could be the combination of two or more rules and the rules may change over time [11]. Starting from this, we proposed our hybrid model with the assumption that networks are evolving following certain rule or the combination of several rules. By finding the rules, we can improve the prediction accuracy.

A. Hybrid Model

Classic topology based link prediction methods work by calculating similarity between nodes [1], [12]. The way how

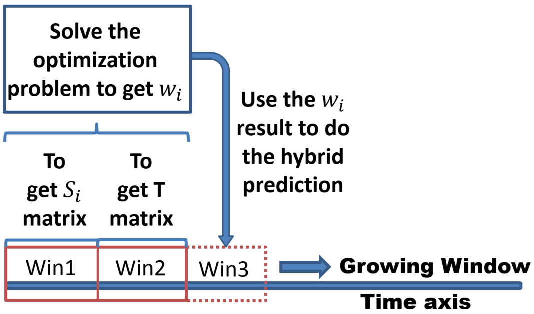
Solve the optimization problem to get w_i
Use the w_i result to do the hybrid prediction
To get S_i matrix
To get T matrix
Win1
Win2
Win3
Growing Window
Time axis

Fig. 1. Hybrid Prediction Model (Growing Window)

the similarity is calculated varies for different methods. For the prediction purposes dataset is split into two sets, the training and the test set, where the training set is used to calculate the similarity score for prediction and the result will be verified using the test set. Our approach differs as we consider link prediction as a time-series problem. As shown in Fig. 1, networks is partitioned into small windows (windows can overlap). We assume the network evolution rule or rules from $Win1$ to $Win2$ remain the same as it is from $Win2$ to $Win3$. Our model is able to work with two scenarios, the growing window and the sliding window. Fig. 1 shows the growing window scenario in which the next action is to grow the window by one step so that we learn the rules from the change from $Win1 \cup Win2$ to $Win3$ and then use it to predict new links in $Win4$. In the sliding window scenario, the model won't memorize the window but will only slide forward. That is, in the next action, we learn new rules from $Win2$ to $Win3$ and use it to predict new links in $Win4$. In this way we enable the method to adapt to the rules that may change over time. To learn the rules, we need to solve the following optimization problem:

$$min(NL - \sum_{i=1}^{k} w_i S_i) \tag{1}$$

Subject to:

$$\sum_{i=1}^{k} w_i = 1; \forall i \in [1,k] : w_i > 0$$

Where NL stands for the new links formed in the window $Win2$ against $Win1$, w_i is the weight assigned to each method, S_i is the similarity score matrix calculated from different selected prediction methods and k is the number of selected prediction methods. The model linearly combines several prediction methods and the rule is the weight vector for each combined prediction method. In our experiment, we use Matlab toolbox CVX [13] to solve the optimization problem.

B. Selected Methods

For hybrid model, we selected eight most widely used topology based link prediction methods as stated in [12].

Common Neighbours method is based on the assumption that the more common neighbours two users have, the higher the probability that a relationship between them will emerge [7], [1], [12]: $|\Gamma(x) \cap \Gamma(y)|$, where $\Gamma(x)$ and $\Gamma(y)$ represents the set of neighbours of node x and node y respectively.

Jaccard's Coefficient is a statistical measure used for comparing similarity of sample sets. In link prediction, all the neighbours of a node are treated as a set and the prediction is done by computing and ranking the similarity of the neighbour set of each node pair. The mathematical expression of this method is as follows [1]: $\left|\frac{\Gamma(x) \bigcap \Gamma(y)}{\Gamma(x) \bigcup \Gamma(y)}\right|$.

Preferential Attachment. Due to the assumption that the node with high degree is more likely to get new links [14], preferential attachment was introduced as a prediction method. This method can be expressed as: $|\Gamma(x)| * |\Gamma(y)|$.

Adamic/Adar Index was initially designed to measure the relation between personal home pages. As shown in equation, the more friends z has, the lower score it will be assigned to. It is calculated as: $\sum_{z \in \Gamma(x) \bigcap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$, where z is a common neighbour of node x and node y.

Resource Allocation method is motivated by the resource allocation dynamics on complex networks[15]. It is very similar to the AA method and the similarity is calculated as: $\sum_{z \in \Gamma(x) \bigcap \Gamma(y)} \frac{1}{\Gamma(z)}$. Comparing with AA method, the difference is insignificant when $\Gamma(z)$ is small, while it is considerable when $\Gamma(z)$ is large.

Cosine Similarity method is based on the dot product of two vectors. It is often used to compare documents in text mining [12]. In network prediction problem, this method is expressed as: $\frac{|\Gamma(x)||\Gamma(y)|}{\|\Gamma(x)\| * \|\Gamma(y)\|}$.

Sørensen Index [16] is designed for comparing the similarity of two samples and originally used to analysis plant sociology. It is defined as: $\frac{2|\Gamma(x) \bigcap \Gamma(y)|}{k_x + k_y}$, where k_x and k_y stands for the degree of node x and node y respectively.

Katz$_\beta$ method takes lengths of all paths between each pair of nodes into consideration [17]. The number of paths between node x and node y with length l (written as $|paths_{xy}^{\langle l \rangle}|$) are calculated and then multiplied by a factor β^l. By summing up all the results for a given two nodes with path length from 1 to ∞, a prediction score for the pair of nodes (x, y) is obtained. $\sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{\langle l \rangle}|$. The parameter β is used to adjust the weight of path with different length. In our experiment, we set $\beta = 0.0005$.

III. Experimental Setting

A. Datasets

We test the hybrid model with two real world network datasets, the Facebook friendship network [18] and the internal email communication network from Wroclaw University of Technology (PWr network), (Table II). In Facebook network, each node represents a user and the link between two nodes means they are friends. For the email communication network, nodes are users and the link represents emails sent between two nodes. Each link in both datasets has a time stamp which records the time when the link was formed. We take both dataset as binary, un-weighted and un-directed networks.

TABLE II
Information about Original Networks

Name	Time Range	Nodes	Links
Facebook	2007/01/01 to 2007-06-30	8,564	33,950
PWr	2008-11-25 to 2009-05-25	14,316	49,950

TABLE I
NETWORKS' CHARACTERISTICS OF GIANT COMPONENTS

Network	**Nodes**	**Links**	**Average Degree**	**Average Shortest Path**	**Diameter**	**Clustering Coefficient**	**Density**
Facebook	7446	23,443	6.297	5.455	15	0.1	0.00042
PWr	6059	27,640	9.124	4.363	20	0.43	0.0015

As shown in table II, there are 14,316 nodes in the PWr Email communication network. However, we find that among these users, only 6,884 users sent email at least once. Rest of the accounts only receive emails. We thus removed all of these nodes with no outgoing activities so that only active users who sent at least one email are kept for the experiment. We also removed from the dataset isolated small cliques as they are not connected with majority of nodes which would bring in noise when perform link prediction. This is achieved by extracting the giant component from both networks. Table I shows the networks' characteristics after the cleansing process.

B. Window Size and Window Step Size

The nodes in the social network are users in real-world social life. Thus, taking into account human social life cycle, we select two window sizes for our experiments – week and month. A week is defined by 7 consecutive days and a month is defined by 28 consecutive days (4 weeks). Another issue is the size of the step that we slide or grow the window by. To address this, we used the method introduced in [19] where authors claim that link prediction accuracy can be increased by choosing window size in a way that the properties of a network within each window are as close as possible to the characteristics of the global network. With considering four characteristics, node degree distribution divergence, the shortest path length distribution divergence, the clustering coefficient divergence and the betweenness centrality divergence introduced in [19], we obtained the optimal step size for both networks. For Facebook, for window of size one month, the optimal step is 14 days and for window of size one week the optimal step is six days. For PWr network, for window of size one month, the optimal step is 28 days and for window of size one week the optimal step is five days. The selected step size applies to both sliding and growing window scenarios.

C. Prediction Accuracy Measure

The prediction performance is measured using recall and precision method. Both precision and recall are numbers between 0 and 1. The higher they are, the more accurate the result. As mentioned in II-A, the prediction methods will calculate the similarity score for each pair of unconnected nodes. We select top N links with the highest similarity score as our predicted new links. Among these links, if only l links are correctly predicted which means they are formed in the next window step, the precision is then define as $Precision = \frac{l}{N}$. Additionally, recall is defined as $Recall = \frac{l}{M}$ where M is the number of all links that should be predicted. In other words recall tells us how many relevant links are predicted and precision how many of the predicted links are relevant.

IV. RESULTS AND DISCUSSION

For each dataset, we run our model for four different scenarios: (i) weekly growing window, (ii) weekly sliding window, (iii) monthly growing window, and (iv) monthly sliding window. 'Weekly' and 'monthly' reffer to the size of the window, i.e. one week and one month respectively and 'sliding' and 'growing' reffer to the methods of selecting the next time windows. The four sub-charts in Fig. 2 - Fig. 5 depict the prediction precision/recall of eight selected prediction methods as well as our hybrid model. The sub-charts (b), (c) and (d) in each figure depict the prediction precision/recall results when we set N as the number of links we would like to predict. N is an arbitrary number between 0 and average number of newly formed links between window steps. The average number of new links is shown in Table III. To make it easier to compare the result between different scenarios and networks, we choose N as $100, 500, 1000$ for both datasets in the scenario of monthly growing and Sliding Windows experiment setting. For weekly growing and sliding windows experiment, we select N as $50, 100, 500$ for both dataset. The (a) original sub-chart depicts the experiment results if we assume that there is the same number of new links formed in the next time step as in the previous one. Because of the limited space we present only results for monthly sliding window scenario. Rest of the results are averaged and presented in Table IV - Table VII. Conducted experiments revealed that both recall and precision of the hybrid model are higher or equal to the highest precision and recall obtained from the eight selected prediction methods separately. We can also observe that the prediction precision as well as recall trends of hybrid model are similar to those of other methods. That is to say if other methods perform well (or poor) in one window step, our hybrid model performs well (or poor) too. This should be expected as the hybrid model is a combination of other methods. It cannot predict new links other than the links predicted by combined methods.

TABLE III
AVERAGE NEW LINKS

	Facebook	PWr
Average Weekly Growing Window New Links	784	732
Average Weekly Sliding Window New Links	784	1003
Average Monthly Growing Window New Links	1815	3763
Average Monthly Sliding Window Min New Links	1815	4142

Facebook Friendship Network.

Precision and recall of the prediction results for Facebook network in the monthly sliding window scenario are shown in figures Fig. 2 and Fig. 3. It can be seen that the hybrid model outperforms other models. The same trend holds for other tested scenarios (weekly/monthly sliding and monthly growing). Table IV shows that on average the best precision is for the prediction of Top 100 links – precision of 0.05 for monthly sliding and 0.063 for monthly growing window. Both the highest precision and average precision drop in the scenario of sliding and growing windows as we increase the number of links we are predicting. Our hybrid model performs better

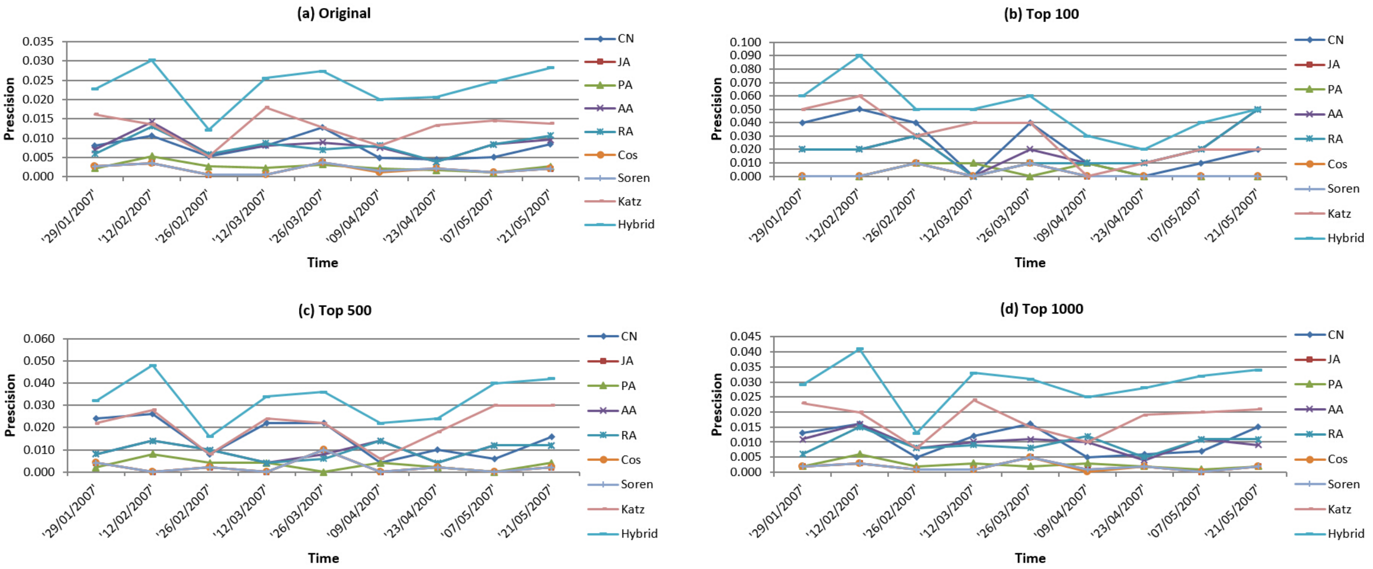

Fig. 2. Facebook Monthly Sliding Window Prediction Precision

TABLE IV
FACEBOOK PREDICTION AVERAGE PRECISION

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0044 (0.0030)	0.0100 (0.0160)	0.0064 (0.0081)	0.0041 (0.0026)	N/A
	Grow		0.0083 (0.0044)	0.0171 (0.0198)	0.0150 (0.0132)	0.0096 (0.0053)	N/A
Monthly	Slide		0.0075 (0.0027)	N/A	0.0233 (0.0183)	0.0153 (0.0080)	0.0106 (0.0045)
	Grow		0.0152 (0.0050)	N/A	0.0289 (0.1500)	0.0218 (0.0060)	0.0174 (0.0050)
Weekly	Slide	JA	0.0025 (0.0017)	0.0029 (0.0088)	0.0018 (0.0047)	0.0022 (0.0022)	N/A
	Grow		0.0005 (0.0009)	0.0000 (0.0000)	0.0004 (0.0019)	0.0004 (0.0012)	N/A
Monthly	Slide		0.0020 (0.0011)	N/A	0.0022 (0.0042)	0.0022 (0.0031)	0.0019 (0.0014)
	Grow		0.0015 (0.0010)	N/A	0.0000 (0.0000)	0.0007 (0.0010)	0.0009 (0.0009)
Weekly	Slide	PA	0.0015 (0.0020)	0.0021 (0.0062)	0.0025 (0.0069)	0.0016 (0.0027)	N/A
	Grow		0.0022 (0.0030)	0.0071 (0.0171)	0.0064 (0.0120)	0.0025 (0.0035)	N/A
Monthly	Slide		0.0026 (0.0011)	N/A	0.0033 (0.0047)	0.0031 (0.0023)	0.0026 (0.0013)
	Grow		0.0039 (0.0036)	N/A	0.0189 (0.0166)	0.0064 (0.0060)	0.0046 (0.0045)
Weekly	Slide	AA	**0.0056** (0.0028)	**0.0114** (0.0155)	**0.0104** (0.0105)	**0.0071** (0.0039)	N/A
	Grow		0.0085 (0.0030)	**0.0221** (0.0240)	**0.0179** (0.0160)	**0.0106** (0.0040)	N/A
Monthly	Slide		0.0082 (0.0027)	N/A	0.0200 (0.0133)	0.0096 (0.0036)	0.0100 (0.0030)
	Grow		0.0129 (0.0040)	N/A	**0.0333** (0.0150)	0.0198 (0.0060)	0.0153 (0.0050)
Weekly	Slide	RA	**0.0056** (0.0028)	**0.0114** (0.0156)	**0.0104** (0.0105)	0.0071 (0.0039)	N/A
	Grow		0.0066 (0.0030)	0.0207 (0.0189)	0.0175 (0.0148)	0.0089 (0.0040)	N/A
Monthly	Slide		0.0080 (0.0026)	N/A	0.0189 (0.0137)	0.0093 (0.0038)	0.0094 (0.0029)
	Grow		0.0099 (0.0020)	N/A	0.0322 (0.0130)	0.0176 (0.0070)	0.0106 (0.0030)
Weekly	Slide	Cos	0.0025 (0.0018)	0.0029 (0.0088)	0.0018 (0.0018)	0.0023 (0.0022)	N/A
	Grow		0.0005 (0.0009)	0.0000 (0.0000)	0.0004 (0.0019)	0.0004 (0.0012)	N/A
Monthly	Slide		0.0020 (0.0012)	N/A	0.0022 (0.0042)	0.0022 (0.0031)	0.0018 (0.0015)
	Grow		0.0013 (0.0009)	N/A	0.0000 (0.0000)	0.0007 (0.0010)	0.0009 (0.0009)
Weekly	Slide	Soren	0.0025 (0.0017)	0.0029 (0.0088)	0.0018 (0.0047)	0.0022 (0.0022)	N/A
	Grow		0.0005 (0.0009)	0.0000 (0.0000)	0.0004 (0.0020)	0.0004 (0.0010)	N/A
Monthly	Slide		0.0020 (0.0011)	N/A	0.0022 (0.0042)	0.0022 (0.003)	0.0019 (0.0014)
	Grow		0.0015 (0.0010)	N/A	0.0000 (0.0000)	0.0007 (0.0010)	0.0009 (0.0009)
Weekly	Slide	Katz	0.0038 (0.0032)	**0.0114** (0.0188)	0.0100 (0.0141)	0.0049 (0.0046)	N/A
	Grow		**0.0094** (0.0051)	0.0186 (0.0226)	0.0154 (0.0145)	0.0103 (0.0055)	N/A
Monthly	Slide		**0.0129** (0.0037)	N/A	**0.0300** (0.0183)	**0.0209** (0.0084)	**0.0178** (0.0053)
	Grow		**0.0158** (0.0050)	N/A	0.0267 (0.0170)	**0.0231** (0.0060)	**0.0189** (0.0070)
Weekly	Slide	Hybrid	*0.0092 (0.0046)*	***0.0235*** *(0.0243)*	*0.0232 (0.0191)*	*0.0126 (0.0065)*	N/A
	Grow		*0.0158 (0.0068)*	***0.0364*** *(0.0321)*	*0.0325 (0.0240)*	*0.0179 (0.0067)*	N/A
Monthly	Slide		*0.0235 (0.0051)*	N/A	***0.0500*** *(0.0189)*	*0.0327 (0.0098)*	*0.0290 (0.0072)*
	Grow		*0.0256 (0.0068)*	N/A	***0.0633*** *(0.0231)*	*0.0382 (0.0120)*	*0.0291 (0.0098)*
Weekly	Slide	Increase	62%	100%	124%	78%	N/A
	Grow		69%	65%	82%	70%	N/A
Monthly	Slide		83%	N/A	67%	56%	66%
	Grow		62%	N/A	90%	65%	54%

when predicting smaller number of links. The optimal number of links that the hybrid model could predict with the highest prediction precision and recall is out of the scope of this study, but it is another interesting topic for future work. For weekly window setting, the highest precision, for both sliding and growing windows, is when Top 50 links is predicted. For the former one it is 0.0235 and for the latter one 0.0364. The standard deviation of the hybrid's model prediction precision is the highest among all the results. It means that the hybrid's model precision fluctuates heavier than other methods but in

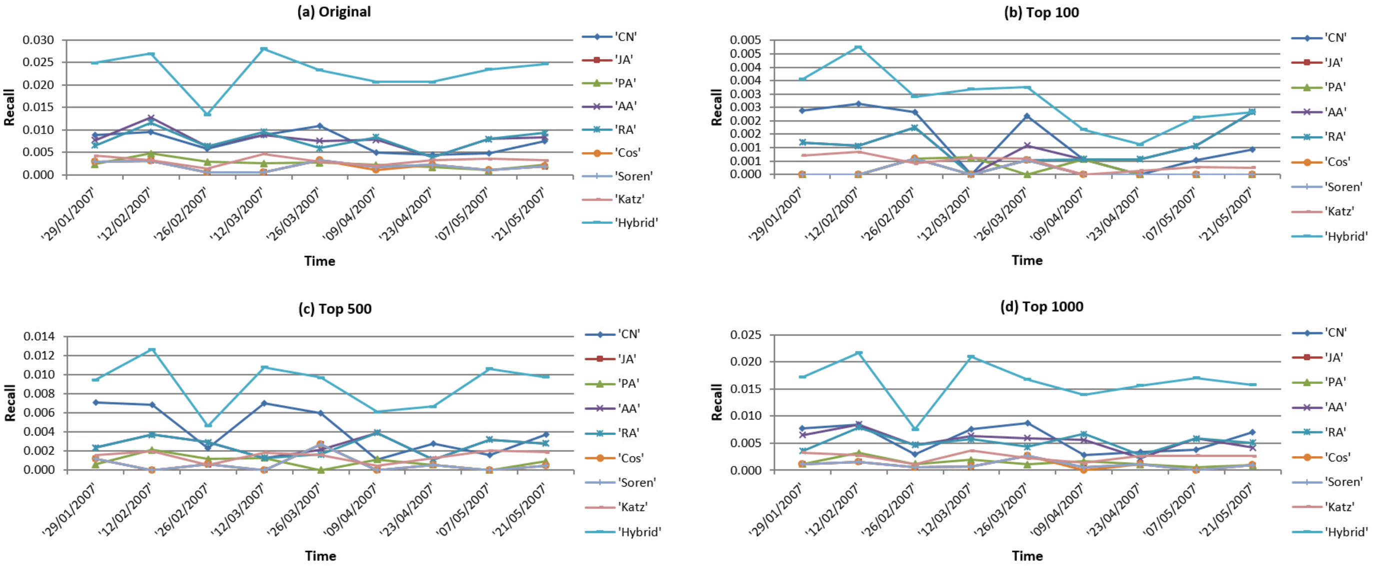

Fig. 3. Facebook Monthly Sliding Window Prediction Recall

TABLE V
AVERAGE RECALL FOR FACEBOOK NETWORK

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0044 (0.0032)	0.0006 (0.0011)	0.0008 (0.0011)	0.0027(0.0017)	N/A
	Grow		0.0082 (0.0043)	0.0011 (0.0012)	0.0019 (0.0016)	0.0062 (0.0034)	N/A
Monthly	Slide		0.0073 (0.0022)	N/A	**0.0013** (0.0010)	**0.0043** (0.0023)	**0.0058** (0.0024)
	Grow		**0.0146** (0.0037)	N/A	0.0016 (0.0010)	**0.0060** (0.0015)	**0.0096** (0.0026)
Weekly	Slide	JA	0.0025 (0.0017)	0.0001 (0.0004)	0.0002 (0.0006)	0.0014 (0.0013)	N/A
	Grow		0.0005 (0.0008)	0.0000 (0.0000)	0.0000 (0.0002)	0.0002 (0.0007)	N/A
Monthly	Slide		0.0019 (0.0010)	N/A	0.0001 (0.0002)	0.0006 (0.0008)	0.0010 (0.0007)
	Grow		0.0014 (0.0010)	N/A	0.0000 (0.0000)	0.0001 (0.0003)	0.0005 (0.0005)
Weekly	Slide	PA	0.0015 (0.0020)	0.0001 (0.0004)	0.0003 (0.0010)	0.0011 (0.0019)	N/A
	Grow		0.0020 (0.0025)	0.0004 (0.0010)	0.0008 (0.0014)	0.0015 (0.0020)	N/A
Monthly	Slide		0.0025 (0.0010)	N/A	0.0002 (0.0003)	0.0009 (0.0006)	0.0014 (0.0007)
	Grow		0.0037 (0.0030)	N/A	0.0010 (0.0009)	0.0017 (0.0016)	0.0024 (0.0024)
Weekly	Slide	AA	**0.0056** (0.0028)	**0.0007** (0.0010)	**0.0013** (0.0013)	**0.0045** (0.0024)	N/A
	Grow		**0.0084** (0.0032)	**0.0014** (0.0015)	**0.0023** (0.0018)	**0.0068** (0.0029)	N/A
Monthly	Slide		**0.0079** (0.0022)	N/A	0.0011 (0.0006)	0.0026 (0.0009)	0.0055 (0.0016)
	Grow		0.0125 (0.0026)	N/A	**0.0019** (0.0009)	0.0055 (0.0017)	0.0084 (0.0026)
Weekly	Slide	RA	**0.0056** (0.0028)	**0.0007** (0.0010)	**0.0013** (0.0012)	**0.0045** (0.0024)	N/A
	Grow		0.0065 (0.0028)	0.0013 (0.0012)	0.0022 (0.0018)	0.0057 (0.0025)	N/A
Monthly	Slide		0.0077 (0.0021)	N/A	0.0010 (0.0007)	0.0025 (0.0010)	0.0052 (0.0015)
	Grow		0.0097 (0.0018)	N/A	0.0018 (0.0007)	0.0048 (0.0015)	0.0058 (0.0013)
Weekly	Slide	Cos	0.0025 (0.0017)	0.0002 (0.0006)	0.0002 (0.0006)	0.0015 (0.0014)	N/A
	Grow		0.0005 (0.0008)	0.0000 (0.0000)	0.0000 (0.0002)	0.0024 (0.0007)	N/A
Monthly	Slide		0.0019 (0.0010)	N/A	0.0001 (0.0002)	0.0006 (0.0008)	0.0010 (0.0008)
	Grow		0.0013 (0.0008)	N/A	0.0000 (0.0000)	0.0002 (0.0003)	0.0005 (0.0005)
Weekly	Slide	Soren	0.0025 (0.0017)	0.0002 (0.0006)	0.0002 (0.0006)	0.0014 (0.0013)	N/A
	Grow		0.0005 (0.0008)	0.0000 (0.0000)	0.0000 (0.0002)	0.0002 (0.0007)	N/A
Monthly	Slide		0.0019 (0.0010)	N/A	0.0001 (0.0002)	0.0006 (0.0008)	0.0010 (0.0007)
	Grow		0.0014 (0.0010)	N/A	0.0000 (0.0000)	0.0002 (0.0003)	0.0005 (0.0005)
Weekly	Slide	Katz	0.0016 (0.0014)	0.0003 (0.0005)	0.0005 (0.0008)	0.0013 (0.0012)	N/A
	Grow		0.0004 (0.0004)	0.0000 (0.0001)	0.0000 (0.0001)	0.0003 (0.0003)	N/A
Monthly	Slide		0.0032 (0.0010)	N/A	0.0004 (0.0002)	0.0015 (0.0006)	0.0025 (0.0008)
	Grow		0.0011 (0.0005)	N/A	0.0001 (0.0000)	0.0005 (0.0002)	0.0008 (0.0004)
Weekly	Slide	Hybrid	***0.0091*** *(0.0047)*	*0.0015 (0.0016)*	*0.0030 (0.0026)*	*0.0080 (0.0043)*	N/A
	Grow		***0.0155*** *(0.0061)*	*0.0022 (0.0019)*	*0.0041 (0.0029)*	*0.0114 (0.0039)*	N/A
Monthly	Slide		***0.0129*** *(0.0041)*	N/A	*0.0028 (0.0010)*	*0.0090 (0.0024)*	*0.0123 (0.0039)*
	Grow		***0.0247*** *(0.0045)*	N/A	*0.0035 (0.0013)*	*0.0105 (0.0030)*	*0.0160 (0.0049)*
Weekly	Slide	Increase	63%	114%	131%	78%	N/A
	Grow		85%	57%	78%	68%	N/A
Monthly	Slide		63%	N/A	115%	109%	112%
	Grow		69%	N/A	84%	75%	67%

the same time they are always above or equal to other results. The last row in Table IV states the improvement rate of our hybrid model over the best performed single prediction method (in bold font) among selected 8 methods. We can see that the hybrid model outperforms other methods by at least 54% and in some cases the improvement rate could be as high as 124%. We can also observe that, for monthly and weekly window setting, the hybrid model performs better in growing window scenario than in the sliding window one. This is due to the fact that in the growing window scenario, the network topology information is aggregated so that the network information is richer in comparison to that in the sliding window scenario.

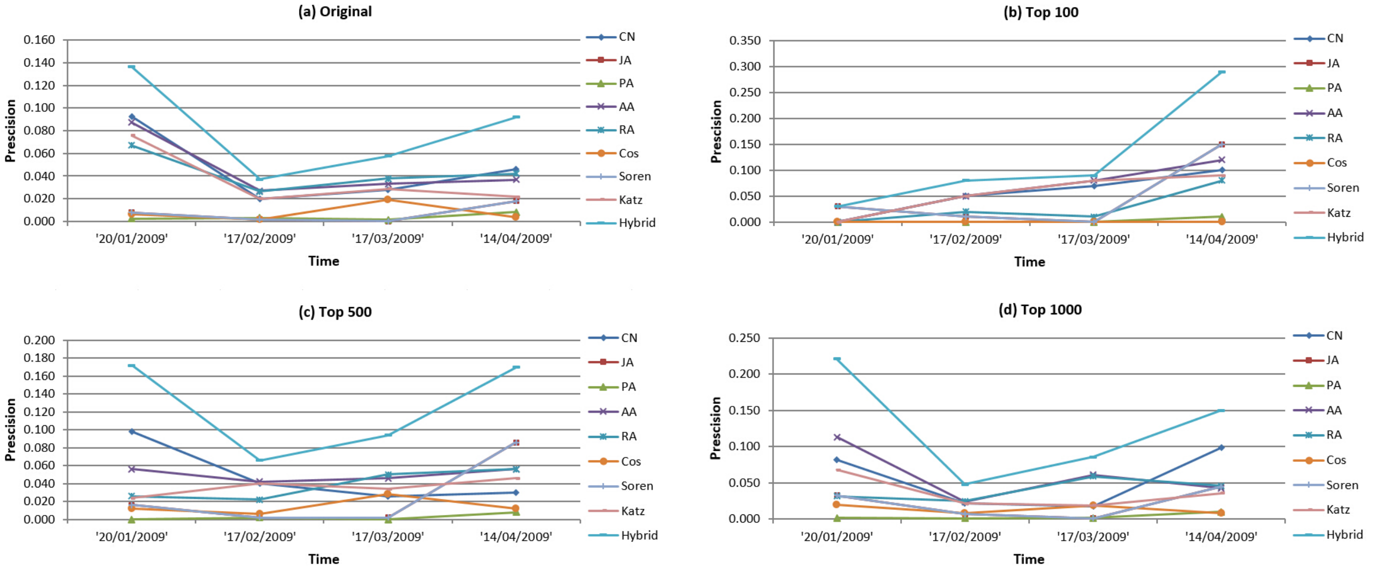

Fig. 4. PWr Monthly Sliding Window Prediction Precision

The richer information helps the model to achieve better prediction result. Similarly, one may think that as window grows, the network topology information gets richer so that the prediction precision should be getting better and better. However, we do not observe a significant increase of precision as window grows for both weekly and monthly experimental settings. When we look at the recall average results (Table V) we can notice that, similarly to precision, the best results are obtained for hybrid model. Moreover, regardless which scenario we consider, the best recall is in a situation when we predict the same number of links as the number of connections formed in the previous time step.

PWr Email Communication Network.
Fig. 4 and Fig. 5 show the prediction precision and recall for PWr network. Similarly to results on the Facebook network, the hybrid model always gives the best prediction outcomes. In the monthly experimental setting, the highest average precision is obtained when Top 500 links is predicted for growing window scenario (with average precision 0.084) and when Top 1000 links is predicted for sliding window scenario (with precision 0.1263). The highest average precision for weekly window setting for growing and sliding scenarios is observed for Top 500 and Top 50 cases with precision of 0.0256 and 0.1406 respectively. In growing window scenario for both weekly and monthly experiment settings, we can observe that precision is high at the beginning and then as window grows precision drop is noticeable. This is very different from that of Facebook prediction results in which we do not find obvious increase and decrease trend. As shown in Table VI, on average, the sliding window results are better than the growing window result. The main reason behind this phenomenon is that if there is no reply for an email then the link might not be valid in the future as the proper relationship has not been formed. So if we simply grow the window, the links formed long time ago, which are no longer valid, have negative effect on the prediction result. The accumulation of this unwanted effect makes the prediction result very poor. The standard deviation of hybrid's model prediction precision in PWr network is similar to that in the Facebook network experiment. The hybrid's model prediction precision is always the best and the standard deviation is larger than other methods as well. The improvement of hybrid model over the best precision result among the 8 selected methods is at least 33% and could be as high as 159%. Results for the recall measure for PWr network follow the same trends as for Facebook network (Table VII). As both, precision and recall, for both analysed networks, are higher than results for other methods, it can suggest that regardless the dataset the hybrid model will remain the best model out of the analysed ones.

V. Conclusions and Future Work

In this work, we claim that online social networks evolve following certain rules that may change over time. Based on this, we introduced a new hybrid link prediction model which was tested on two real world online social networks of different types, the contact–based Facebook network and activity–based email network. The results of the experiments show that the prediction precision and recall of hybrid model are higher than of any of the other tested methods. Although the model outperforms all of the selected methods, it still has a limit. As the model is a combination of selected methods, its prediction results heavily relies on the results of selected methods. It also explains why the changes in the precision/recall levels of the hybrid model always follow the changes in the precision/recall of other well performing methods.

The prediction precision results of the two networks are different. For Facebook network, the average prediction precision of hybrid model with growing window scenario are better than that with sliding window scenario whereas for PWr network the results of hybrid model in the sliding window scenario are much better than those in the growing window scenario. In email communication network, links are formed by sending emails between two nodes. These links are only

TABLE VI
AVERAGE PRECISION FOR PWR NETWORK

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0273 (0.0388)	**0.0735** (0.1043)	0.0703 (0.1069)	0.0423 (0.0547)	N/A
	Grow		0.0136 (0.0219)	0.0029 (0.0099)	0.0029 (0.0062)	**0.0154** (0.0239)	N/A
Monthly	Slide		**0.0466** (0.0281)	N/A	0.0550 (0.0364)	**0.0485** (0.0290)	0.0553 (0.0358)
	Grow		0.0309 (0.0346)	N/A	0.0025 (0.0043)	0.0495 (0.0618)	**0.0480** (0.0630)
Weekly	Slide	JA	0.0040 (0.0059)	0.0165 (0.0438)	0.0124 (0.0311)	0.0068 (0.0131)	N/A
	Grow		0.0007 (0.0014)	0.0041 (0.0117)	0.0024 (0.0060)	0.0012 (0.0021)	N/A
Monthly	Slide		0.0070 (0.0069)	N/A	0.0475 (0.0602)	0.0265 (0.0348)	0.0215 (0.0181)
	Grow		0.0032 (0.0015)	N/A	0.0075 (0.0083)	0.0075 (0.0107)	0.0043 (0.0062)
Weekly	Slide	PA	0.0028 (0.0069)	0.0047 (0.0119)	0.0041 (0.0109)	0.0038 (0.0123)	N/A
	Grow		0.0014 (0.0042)	0.0006 (0.0034)	0.0006 (0.0024)	0.0014 (0.0038)	N/A
Monthly	Slide		0.0038 (0.0026)	N/A	0.0025 (0.0043)	0.0025 (0.0033)	0.0038 (0.0036)
	Grow		0.0010 (0.0004)	N/A	0.0000 (0.0000)	0.0020 (0.0014)	0.0018 (0.0011)
Weekly	Slide	AA	**0.0350** (0.0365)	0.0424 (0.0691)	0.0474 (0.0608)	**0.0439** (0.0465)	N/A
	Grow		0.0107 (0.0151)	0.0018 (0.0075)	0.0024 (0.0055)	0.0124 (0.0191)	N/A
Monthly	Slide		0.0460 (0.0241)	N/A	**0.0625** (0.0438)	0.0500 (0.0062)	**0.0603** (0.0331)
	Grow		**0.0338** (0.0356)	N/A	0.0025 (0.0043)	0.0350 (0.0410)	0.0453 (0.0543)
Weekly	Slide	RA	0.0296 (0.0201)	0.0276 (0.0333)	0.0241 (0.0301)	0.0336 (0.0296)	N/A
	Grow		0.0064 (0.0079)	0.0012 (0.0047)	0.0035 (0.0080)	0.0051 (0.0078)	N/A
Monthly	Slide		0.0434 (0.0148)	N/A	0.0275 (0.0311)	0.0385 (0.0147)	0.0405 (0.0131)
	Grow		0.0248 (0.0190)	N/A	0.0000 (0.0000)	0.0105 (0.0078)	0.0135 (0.0097)
Weekly	Slide	Cos	0.0070 (0.0093)	0.0176 (0.0349)	0.0121 (0.0240)	0.0092 (0.0106)	N/A
	Grow		0.0026 (0.0052)	**0.0065** (0.0226)	**0.0082** (0.0232)	0.0034 (0.0071)	N/A
Monthly	Slide		0.0077 (0.0067)	N/A	0.0000 (0.0000)	0.0145 (0.0082)	0.0138 (0.0058)
	Grow		0.0061 (0.0033)	N/A	**0.0300** (0.0520)	0.0185 (0.0175)	0.0105 (0.0104)
Weekly	Slide	Soren	0.0040 (0.0059)	0.0165 (0.0438)	0.0124 (0.0311)	0.0068 (0.0131)	N/A
	Grow		0.0007 (0.0014)	0.0041 (0.0117)	0.0024 (0.0060)	0.0012 (0.0021)	N/A
Monthly	Slide		0.0070 (0.0069)	N/A	0.0475 (0.0602)	0.0265 (0.0348)	0.0215 (0.0181)
	Grow		0.0032 (0.0015)	N/A	0.0075 (0.0083)	0.0075 (0.0107)	0.0043 (0.0062)
Weekly	Slide	Katz	0.0240 (0.0326)	0.0771 (0.1025)	**0.0756** (0.1118)	0.0435 (0.0594)	N/A
	Grow		**0.0136** (0.0205)	0.0029 (0.0099)	0.0029 (0.0062)	0.0139 (0.0218)	N/A
Monthly	Slide		0.0366 (0.0229)	N/A	0.0550 (0.0350)	0.0360 (0.0081)	0.0363 (0.0194)
	Grow		0.0330 (0.0379)	N/A	0.0025 (0.0043)	**0.0395** (0.0493)	0.0413 (0.0514)
Weekly	Slide	Hybrid	*0.0554 (0.0455)*	***0.1406*** *(0.1278)*	*0.1256 (0.1326)*	*0.0814 (0.0696)*	N/A
	Grow		*0.0241 (0.0323)*	*0.0141 (0.0345)*	*0.0162 (0.0089)*	***0.0256*** *(0.0347)*	N/A
Monthly	Slide		*0.0808 (0.0376)*	N/A	*0.1225 (0.0993)*	*0.1255 (0.0466)*	***0.1263*** *(0.0657)*
	Grow		*0.0549 (0.0456)*	N/A	*0.0400 (0.0636)*	***0.0840*** *(0.0903)*	*0.0790 (0.0907)*
Weekly	Slide	Increase	58%	91%	66%	85%	N/A
	Grow		77%	118%	97%	67%	N/A
Monthly	Slide		73%	N/A	96%	159%	110%
	Grow		63%	N/A	33%	113%	65%

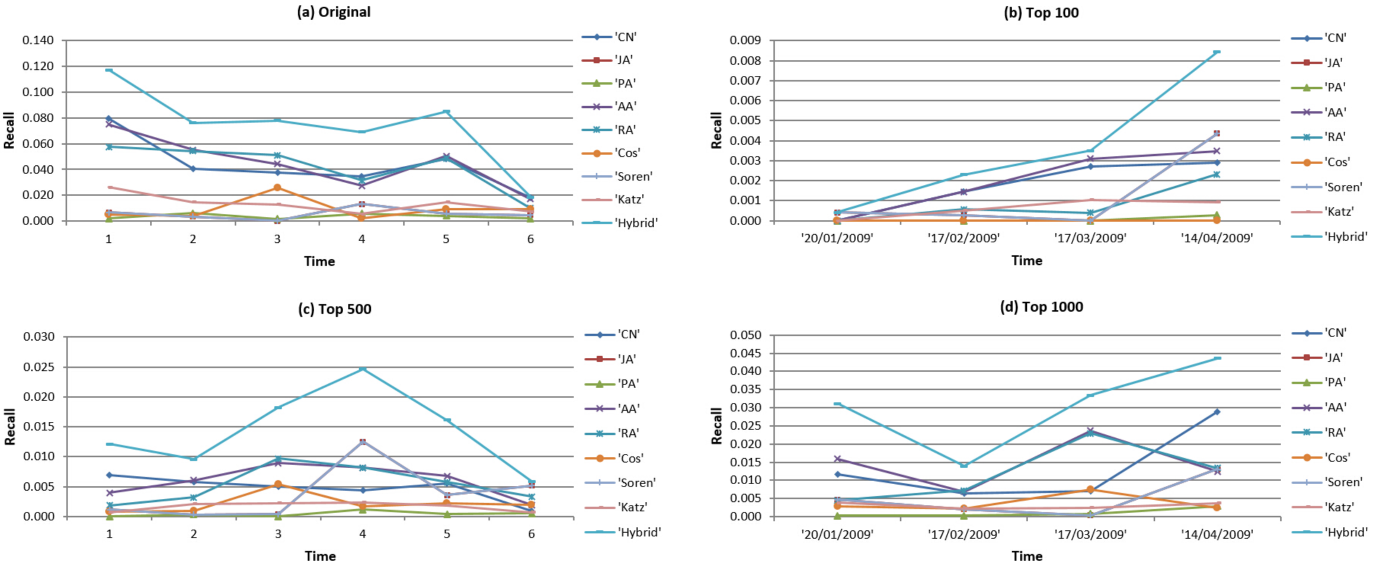

Fig. 5. PWr Monthly Sliding Window Prediction Recall

valid for a few days and thus growing window approach does not help in the link prediction task. Links that are only valid for short period of time introduce a lot of noise in the long term prediction. However, in the case of Facebook network, link represents a friendship and it lasts much longer than email network link. For this type of networks, growing window prediction approach performs better than sliding windows prediction. Taking the above into account, we can conclude that the networks are evolving in different ways. Selecting the proper experiment scenario (i.e. sliding window or growing

TABLE VII
AVERAGE RECALL FOR PWR NETWORK

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0273 (0.0395)	**0.0042** (0.0086)	**0.0084** (0.0203)	0.0199(0.0277)	N/A
	Grow		**0.0165** (0.0278)	0.0001 (0.0004)	0.0009 (0.0042)	**0.0082** (0.0102)	N/A
Monthly	Slide		0.0481 (0.0181)	N/A	0.0018 (0.0012)	0.0055 (0.0009)	0.0134 (0.0091)
	Grow		0.0317 (0.0293)	N/A	0.0000 (0.0000)	**0.0047** (0.0041)	**0.0089** (0.0083)
Weekly	Slide	JA	0.0042 (0.0078)	0.0007 (0.0017)	0.0012 (0.0027)	0.0037 (0.0066)	N/A
	Grow		0.0006 (0.0010)	**0.0003** (0.0006)	0.0003 (0.0007)	0.0006 (0.0010)	N/A
Monthly	Slide		0.0061 (0.0048)	N/A	0.0013 (0.0018)	0.0036 (0.0052)	0.0050 (0.0048)
	Grow		0.0036 (0.0018)	N/A	0.0002 (0.0002)	0.0006(0.0008))	0.0007 (0.0009)
Weekly	Slide	PA	0.0023 (0.0051)	0.0002 (0.0006)	0.0004 (0.0010)	0.0017 (0.0050)	N/A
	Grow		0.0013 (0.0032)	0.0000 (0.0002)	0.0000 (0.0002)	0.0007 (0.0017)	N/A
Monthly	Slide		0.0041 (0.0021)	N/A	0.0000 (0.0000)	0.0004 (0.0005)	0.0011 (0.0011)
	Grow		0.0012 (0.0007)	N/A	0.0000 (0.0000)	0.0003 (0.0003)	0.0004 (0.0002)
Weekly	Slide	AA	**0.0380** (0.0382)	0.0023 (0.0042)	0.0052 (0.0077)	**0.0244** (0.0287)	N/A
	Grow		0.0134 (0.0189)	0.0000 (0.0003)	**0.0010** (0.0042)	0.0065 (0.0088)	N/A
Monthly	Slide		**0.0505** (0.0171)	N/A	**0.0020** (0.0014)	**0.0068** (0.0019)	**0.0148** (0.0061)
	Grow		**0.0366** (0.0301)	N/A	0.0000 (0.0000)	0.0034 (0.0028)	0.0088 (0.0069)
Weekly	Slide	RA	0.0374 (0.0381)	0.0016 (0.0018)	0.0030 (0.0042)	0.0230 (0.0270)	N/A
	Grow		0.0088 (0.0113)	0.0000 (0.0002)	0.0004 (0.0010)	0.0036 (0.0056)	N/A
Monthly	Slide		0.0486 (0.0100)	N/A	0.0008 (0.0009)	0.0057 (0.0033)	0.0120 (0.0070)
	Grow		0.0301 (0.0179)	N/A	0.0000 (0.0000)	0.0013 (0.0003)	0.0032 (0.0008)
Weekly	Slide	Cos	0.0082 (0.0133)	0.0011 (0.0022)	0.0015 (0.0034)	0.0068 (0.0115)	N/A
	Grow		0.0020 (0.0042)	**0.0003** (0.0009)	**0.0010** (0.0027)	0.0021 (0.0041)	N/A
Monthly	Slide		0.0095 (0.0094)	N/A	0.0000 (0.0000)	0.0022 (0.0019)	0.0037 (0.0021)
	Grow		0.0073 (0.0050)	N/A	**0.0004** (0.0007)	0.0022 (0.0011)	0.0024 (0.0014)
Weekly	Slide	Soren	0.0042 (0.0078)	0.0007 (0.0017)	0.0012 (0.0027)	0.0037 (0.0066)	N/A
	Grow		0.0006 (0.0011)	**0.0003** (0.0007)	0.0003 (0.0007)	0.0006 (0.0009)	N/A
Monthly	Slide		0.0061 (0.0048)	N/A	0.0013 (0.0018)	0.0036 (0.0052)	0.0050 (0.0049)
	Grow		0.0036 (0.0018)	N/A	0.0002 (0.0002)	0.0006 (0.0008)	0.0007 (0.0008)
Weekly	Slide	Katz	0.0059 (0.0075)	0.0010 (0.0015)	0.0019 (0.0033)	0.0049 (0.0062)	N/A
	Grow		0.0008 (0.0013)	0.0000 (0.0000)	0.0000 (0.0000)	0.0004 (0.0006)	N/A
Monthly	Slide		0.0149 (0.0074)	N/A	0.0006 (0.0004)	0.0018 (0.0007)	0.0031 (0.0007)
	Grow		0.0049 (0.0068)	N/A	0.0000 (0.0000)	0.0005 (0.0006)	0.0011 (0.0015)
Weekly	Slide	Hybrid	***0.0597*** *(0.0488)*	*0.0078 (0.0095)*	*0.0140 (0.0213)*	*0.0443 (0.0329)*	N/A
	Grow		***0.0269*** *(0.0365)*	*0.0006 (0.0014)*	*0.0025 (0.0049)*	*0.0141 (0.0144)*	N/A
Monthly	Slide		***0.0849*** *(0.0188)*	N/A	*0.0036 (0.0029)*	*0.0162 (0.0059)*	*0.0305 (0.0107)*
	Grow		***0.0615*** *(0.0371)*	N/A	*0.0006 (0.0009)*	*0.0087 (0.0052)*	*0.0161 (0.0110)*
Weekly	Slide	Increase	57%	86%	67%	82%	N/A
	Grow		63%	100%	150%	72%	N/A
Monthly	Slide		68%	N/A	80%	138%	106%
	Grow		68%	N/A	50%	85%	81%

window) helps to improve the prediction accuracy of our hybrid model. Although, when looking at the improvement rate of hybrid model over others we see significant improvement, the absolute prediction precision and recall values remain low. In our experiment, we only applied the eight well-known prediction methods. However, in the future we plan to introduce community information into proposed hybrid model as well as information about nodes and edges characteristics.

REFERENCES

[1] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the twelfth international conference on Information and knowledge management*, ser. CIKM '03. New York, NY, USA: ACM, 2003, pp. 556–559.

[2] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, "Epidemic thresholds in real networks," *ACM Trans. Inf. Syst. Secur.*, vol. 10, no. 4, pp. 1:1–1:26, Jan. 2008.

[3] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, 2001.

[4] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic spreading in real networks: An eigenvalue viewpoint," in *In SRDS*, 2003, pp. 25–34.

[5] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, ser. JCDL '05. New York, NY, USA: ACM, 2005, pp. 141–142.

[6] F. Molnar, "Link Prediction Analysis in the Wikipedia Collaboration Graph," 2011. [Online]. Available: http://assassin.cs.rpi.edu/~magdon/courses/casp/projects/Molnar.pdf

[7] W. Cukierski, B. Hamner, and B. Yang, "Graph-based features for supervised link prediction," in *The 2011 International Joint Conference on Neural Networks*, 2011, pp. 1237–1244.

[8] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *Proceedings of the Third IEEE International Conference on Social Computing*, 2011, pp. 73–80.

[9] K. Juszczyszyn, K. Musial, and M. Budka, "Link prediction based on subgraph evolution in dynamic social networks," in *Third IEEE International Conference on Social Computing*, 2011, pp. 27–34.

[10] F. Gao, K. Musial, C. Cooper, and S. Tsoka, "Link prediction methods and their accuracy for different social networks and network metrics," *Scientific Programming*, vol. 2015, pp. 172 879:1–172 879:13, 2015. [Online]. Available: http://dx.doi.org/10.1155/2015/172879

[11] K. Musial, M. Budka, and K. Juszczyszyn, "Creation and growth of online social network - how do social networks evolve?" *World Wide Web*, 2012.

[12] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A Statistical Mechanics and its Applications*, vol. 390, pp. 1150–1170, 2011.

[13] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[14] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, vol. 64, p. 025102, Jul 2001.

[15] Q. Ou, Y. D. Jin, T. Zhou, B. H. Wang, and B. Q. Yin, "Power-law strength-degree correlation from resource-allocation dynamics on weighted networks," *Phys. Rev. E*, vol. 75, p. 021102, 2007.

[16] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Biol. Skr.*, vol. 5, pp.

1–34, 1948.

[17] L. Katz, “A new status index derived from sociometric analysis,” *Psychometrika*, vol. 18, pp. 39–43, 1953. [Online]. Available: http://dx.doi.org/10.1007/BF02289026

[18] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, “On the evolution of user interaction in facebook,” in *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN’09)*, August 2009.

[19] M. Budka, K. Musial, and K. Juszczyszyn, “Predicting the evolution of social networks: Optimal time window size for increased accuracy,” in *The Fourth IEEE International Conference on Social Computing*, 2012, pp. 21–30.