Exploring Activity Features in Predicting Social Network Evolution

Shu Huang
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802, USA
Email: shuang@ist.psu.edu

Dongwon Lee
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802, USA
Email: dongwon@psu.edu

***Abstract*—In this paper, we present a novel approach to incorporate the activity features in measuring the influence of member activities on the social network evolution. Conventional methods analyze social networks and make predictions based on all cumulative members and activities. However, since inactive members do not contribute to the network growth, including them in analysis can lead to less accurate results. Based on this observation, we propose to focus on the active population and explore the influence of member activities. We present a model that can incorporate various activity features and predict the evolution of social activities. At the same time, an algorithm is adopted to select the most influential activity features. The experiments on two different types of social network show that the activity features can predict the evolution of the social activity accurately and our algorithm is effective to select the most influential features. In addition, we find that the most significant activity features to determine the network evolution vary among different types of social network.**

***Keywords*-Social Activity Evolution, Feature Selection**

I. INTRODUCTION

The emergence of digital libraries and online communities provides a lot of resources that enable in-depth research in social network analysis. Exploring the network evolution provides insights to the change of relationships between people. Existing studies involve different kinds of networks, such as blogger network, email communities, and academic co-authorship network. Investigating what factors are influential to the social activity is critical to understand the evolution of social networks.

In business such as advertising, studies on the network evolution enable people to forecast the future network status and thus determine the corresponding marketing strategies in advance. For instance, if members who have frequent activities (i.e. active members) keep increasing, it is probably worthwhile to invest in advertisements and promotions over this growing network. Within a network, active members are more important than inactive ones, because the former reveals the real value of the current community and contribute more to the future status of the network. The applications in real marketing scenarios make it necessary to focus on active members and investigate how their activities affect the network evolution.

This work was in part supported by NSF DUE-0817376 and DUE-0937891 awards

Different models are proposed to address the social network analysis in past years. Many studies aim to capture microscopic structures and simulate the network growth with additions of new vertices [1][2], but they are mostly based on the cumulative infrastructure and do not emphasize the current member activities. Recent studies pay more attention to analyze the structural features and their evolution over time [3]. However, the influence of these features on network evolution is only evaluated at individual level. In [4], while qualitative evolution patterns of co-authorship network are summarized, the impact of specific features is not addressed.

Based on these observations, we propose to focus on the active population of a social network and explore the impact of member activities on its evolution. We propose a model that can incorporate various activity features and predict the network evolution quantitatively. Furthermore, we investigate the most significant activity features in determining the network evolution, and then make use of them in prediction. Finally, we show by experiment that our method is effective to select the most valuable features and improve the accuracy of prediction.

The contributions of this paper are summarized as follows:

- We propose a novel approach to measure the social network evolution based on active members and investigate the influence of member activities upon that.
- We incorporate activity features as well as their temporal effects, and provide insights into the relationship between them and the macroscopic evolution of a network.
- We introduce a method to find out the most significant activity features and furthermore predict the active network evolution with them.

In the experiment, the Citeseer co-authorship network and Facebook online network are used for evaluation. The selected significant activity features are different over different types of social network. On the facebook data, it is found that the number of current active members and features relevant to the number of edges could be the most informative factors to predict the online social network evolution. On the Citeseer data, however, it is observed that

the collaboration between members could be an important indicator to explain the evolving pattern of the co-authorship network.

The rest of this paper is organized as follows. Section 2 discusses the related work. The preliminaries and definitions are shown in Section 3. In Section 4, we present details of the methods for the feature selection and the network evolution prediction. After that, we show the experiment results in Section 5. The conclusions are drawn in Section 6.

II. Related Work

Large scale social network evolution has been studied and explored intensively in recent years. Generally, existing work can be categorized into two classes: microscopic evolution modeling and time-evolving structure analysis.

The microscopic evolution modeling focuses on the attachment of new edges and the arrival of new vertices. Preferential attachment of new vertices is proposed to follow the power-law degree distribution in [5][6]. Another microscopic evolution model was proposed with nodes arriving at a prespecified rate and selecting their lifetimes [7]. These models address the network evolution with a micro scope, but they overlook the macroscopic structural predictors.

Time-evolving structure analysis pays more attention to the structural feature measures and their evolution patterns [8]. Different behavior scaling in degree distribution is analyzed on online social networks [9]. The forest fire model is presented to explain the densification and shrinking diameters over time [10]. Another analysis on the co-authorship network explores the community growth with topic change[11]. However, in these studies, the influence of structural properties is investigated at individual level and their measures of growth are based on the cumulative additions of new vertices.

Many existing works either focus on the microscopic infrastructure or ignore activity properties when simulating the network evolution. Different from them, we address the social activity from a macro scope, and study the impact of activity features on it. Incorporating these features, we propose a method to predict the evolution and find out the most influential features in the process.

III. Preliminaries

An underlying intuition of social network evolution is that the activities of members have effects on it and may determine its future status. In most existing studies, once an individual joins a community, her membership is considered valid regardless whether she has activities or not in future. However, this is not consistent with the fact that social activities may also decrease over time. Therefore, we propose that if an individual does not have any activity during a time period, she may not contribute to the community evolution and thus should not be considered in the evolution study over that period.

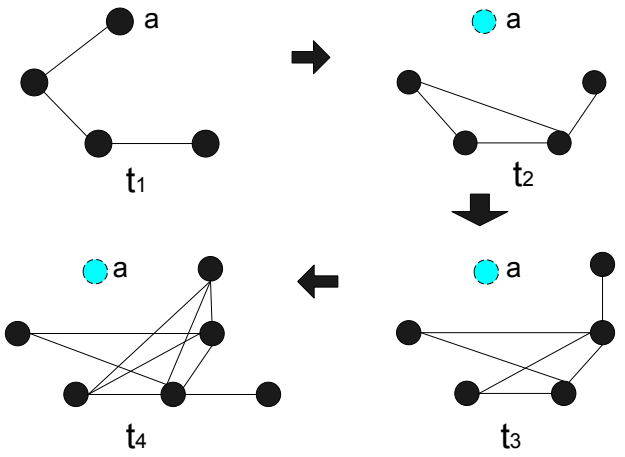

Figure 1. An example of Facebook social network evolution

Figure1 shows an example of a group evolution in the Facebook wall-posting social network over a period of four time steps. At time t_1, four nodes represent four members and they have four posts on the walls of each other. At time t_2, member a does not post any more, but others continue to be active and a new member joins in. At time t_3, more posts appear between other members and a second new member joins in, while a is still inactive. At time t_4, existing members are more active and the community keeps expanding, but a does not contribute into neither the activity nor the community growth.

From Figure 1, we can see that not all members are consistently active after joining a community. The more active members are, the more new members the community is likely to attract. Additionally, not all members in a social network contribute to the growth of the social activities equally. Only existing active members are involved in the activities. Therefore the inactive members should not be included when evaluating the influence of members on the social activity evolution. Based on these observations, we formalize a concept of *interval-wise social activity*(ISA) as follows:

Definition 1(Interval-wise Social Activity): Given a group of individuals M with activities during time interval $[t_0, t_n]$, the interval-wise social activity of M at time interval $[t_i, t_{i+1})$ is represented by $G(t_i) = (V_{t_i}, E_{t_i})$, where V_{t_i} is the vertex set and E_{t_i} is the edge set. Every vertex $v \in V_{t_i}$ corresponds to an individual who is involved in at least k activities on $[t_i, t_{i+1})$, where k is the cut-off threshold. An edge $e =< u, v >, e \in E_{t_i}$ exists between a pair of vertex u and v if and only if u and v have at least k interactions of a particular activity type during the interval $[t_i, t_{i+1})$.

Within an ISA, the membership is no longer permanent.

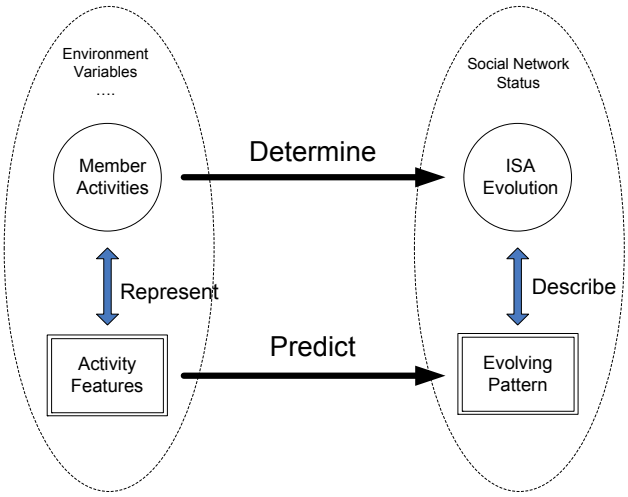

Figure 2. The interaction between member activities and social network evolution

Only those who have activities during $[t_i, t_{i+1})$ are included in $G(t_i)$. By focusing on the macroscopic evolution, let N_t denote the ISA status. The measure of N_t is flexible and may vary according to applications. In our experiment, we use the community size $|V(t_i)|$ as the status measure, i.e. $N_{t_i} = |V(t_i)|$. To describe the evolution of an ISA, we define another concept of *evolving pattern* as follows:

Definition 2(Evolving Pattern) : The evolving pattern is $D = \mathrm{d}N/\mathrm{d}t$. In a discrete format, $D_{t_i} = (N_{t_{i+1}} - N_{t_i})/\Delta t$, where Δt is a fixed time interval. Based on different values of D_{t_i}, evolving pattern has two labels: growing and shrinking. Growing is when $D_{t_i} >= 0$, and shrinking is when $D_{t_i} < 0$.

In this study, we only focus on the labels of the evolving pattern, and use binary marks L to represent the two labels. The value of L is 1, if the ISA is growing in the next time interval, and 0 if shrinking.

The evolving pattern measures the evolution of ISA from the macro scope, instead of the micro scope. Taking no account of random factors and environment variables, it is the member interaction in $G(t)$ that attracts new members and determines the future ISA status. When studying the influence of member activities on the evolving pattern, the activity features may serve as a good summary of the member activity. Figure 2 illustrates the interactive relationship between the ISA evolution and current member activities as well as their alternative representation. The description of member activities are quantified by investigating and integrating different activity features. At the same time, the ISA evolution is measured by temporal evolving patterns.

By using activity features, the problem of finding the impact of the member activity can be transformed to measuring the influence of activity features on the evolving pattern. Therefore, we extract activity features from ISA and employ a model that can select the most significant features in explaining and predicting the evolving pattern.

In evolution prediction, the full feature set may not produce the highest prediction accuracy. One explanation could be that with more variables, the correlation between these variables may introduce much more bias into the result. This leads to the overall decrease of the prediction accuracy, which is called the "overfitting" problem. Therefore, it is necessary to select the most efficient set of activity features.

Given the overall analysis, the goal of our study is to determine the fitting model and select as few activity features as possible that produce the best prediction accuracy. In the next section, we introduce the activity features extracted from ISA and apply a regression model to predict the evolving pattern with them. At the same time, another method is adopted to select the most significant features in explaining and predicting the evolving pattern.

IV. Activity Features Selection and Evolving Pattern Prediction

In this section, we present the details of activity features extracted from two types of social networks. After that, we introduce an approach to select the most significant features and a method to predict the evolving pattern with them.

A. Feature Extraction

Although different social networks have different infrastructures, they usually share a lot of common activity features. Two representative social networks are explored in our study: Citeseer co-authorship network and Facebook wall-posting social network. In the dataset of Citeseer, the co-authorship is considered as the interactive activity between members. In Facebook dataset, the interaction between individuals is posting on each other's wall. The edges in Citeseer and Facebook networks are both considered unweighted, which means an edge exist between two individuals as long as they have at least one interaction at the time.

To make the measure more flexible, the value of evolving pattern is not used directly in selecting activity features selection. Instead, we use the labels of the evolving pattern and their binary marks on both Citeseer and Facebook datasets.

To measure ISA and the member activity therein, the activity records of a social network are partitioned on successive and identical time intervals. Within each interval, we extract some activity features that are frequently measured in social network analysis [12]. The feature set includes not only the characteristics related exclusively to the social network structure, but also those indicating the activity level of the members.

Tables I and II summarize all the activity features generated from Citeseer and Facebook dataset respectively. The activity features extracted for two social networks are not exactly the same because of their different infrastructures. To

Table I
ACTIVITY FEATURES OF CITESEER CO-AUTHORSHIP NETWORK

Feature Notation	Activity Feature Description
$\|V_t\|$	active member population in $G(t)$
CV_t	cumulative active member population by time t
ΔV_t	difference between V_t and V_{t-1}
P_t	number of all publications in $G(t)$
CP_t	cumulative number of all publication by time t
ΔP	difference between P_t and P_{t-1}
$\|E_t\|$	number of edges in $G(t)$
CE_t	cumulative number of edges by time t
ΔE_t	difference between E_t and E_{t-1}
C_t	number of all collaborations in $G(t)$
ΔC_t	difference between C_t and C_{t-1}
CMC_t	cumulative number of collaboration by time t
AR_t	average number of coauthors per person in $G(t)$
ΔAR_t	difference between AR_t and AR_{t-1}
CC_t	average clustering coefficient in $G(t)$
ΔCC_t	difference between CC_t and CC_{t-1}
AL_t	average length of the shortest pathes in $G(t)$
D_t	diameter across all vertices of $G(t)$

Table II
ACTIVITY FEATURES OF FACEBOOK ONLINE WALL-POSTING SOCIAL NETWORK

Feature Notation	Activity Feature Description
$\|V_t\|$	active member population in $G(t)$
CV_t	cumulative active member population by time t
ΔV_t	difference between V_t and V_{t-1}
P_t	number of all posts in $G(t)$
CP_t	cumulative number of all posts by time t
ΔP	difference between P_t and P_{t-1}
$\|E_t\|$	number of edges in $G(t)$
CE_t	cumulative number of edges by time t
ΔE_t	difference between E_t and E_{t-1}
AI_t	average number of interaction per person in $G(t)$
ΔAI_t	The difference between AI_t and AI_{t-1}
AP_t	average number of posts for each person in $G(t)$
ΔAP_t	The difference between AP_t and AP_{t-1}
CC_t	average clustering coefficient in $G(t)$
ΔCC_t	difference between CC_t and CC_{t-1}
AL_t	average length of the shortest pathes in $G(t)$
D_t	diameter across all vertices of $G(t)$

handle very large values, some activity features are rescaled by logarithm in our experiment, e.g. $|V_t|$, P_t, and $|E_t|$.

B. Feature Selection and Evolution Prediction

1) The Shrinkage Method: In selecting variables, using shrinkage methods allows a variable to be partly included in a fitting model. The shrunken coefficient indicates how much information the variable contributes as a factor in the model. The lasso (Least Absolute Shrinkage and Selection Operator) is a widely accepted method[13]. We apply the lasso as a shrinkage method with our fitting model and select the most significant activity features.

Take lasso with a linear fitting model as an illustration. Given a set of input variables $x_1, x_2, ...x_p$ and a response Y, the linear regression fitting model is expressed as:

$$\hat{Y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

The lasso fits the model and estimates the coefficients $\beta = (\beta_0, \beta_1, ...\beta_p)$ by the criterion

$$\hat{\beta} = argmin(\sum (Y - \hat{Y})^2) \quad (1)$$

subject to

$$\sum |\beta_i| \leq s \quad (2)$$

where $s(> 0)$ is a user-specified parameter.

The criterion of the lasso is to minimize the residual sum of squares subject to the constraint (2), where the parameter s is often set moderately small (e.g. $s = 1$). In application, the fitting model is not restricted to be linear regression.

2) Evolving Pattern Prediction: Since evolving patterns are labeled with binary marks, the response of the model is categorical, instead of numerical. Thus, we adopt the logistic regression as the fitting model, which has less assumption on the data and is more robust than the linear model. The results of logistic regression are probabilistic, instead of binary. They enable a flexible optimal boundary to assign the evolving pattern marks.

Suppose $Y_1, Y_2, ..., Y_n$ are independent binary response variables, which denote the evolving pattern marks and take the value of 1 or 0. The activity features are predictor variables represented by $x_1, x_2, ..., x_n$ with $x_i = (1, x_{i1}, x_{i2}, ..., x_{ip})'$, $1 \leq i \leq n$, where p and n specify the number of variables and the size of the dataset, respectively. Define $\pi(t) = \frac{e^t}{1+e^t}$, according to the logistic regression, the distribution function is:

$$P(Y_i = 1|x_i) = \pi(x_i{}'\beta) = \frac{exp(x_i{}'\beta)}{1 + exp(x_i{}'\beta)} \quad (3)$$

where $1 \leq i \leq n$, and $\beta = (\beta_0, \beta_1, ...\beta_p)$ is a (p+1) dimensional vector of coefficients including the intercept. The optimal decision boundary of $P(Y_i|x_i)$ could be determined by achieving the best prediction accuracy on the training set. For simplicity, we take 0.5 as the decision boundary in this study.

3) Activity Features Selection: Adopting the logistic regression as our fitting model, we apply the lasso to it and perform the activity feature selection with an iterative algorithm.

Based on the equation (3), the log-likelihood function of logistic regression is:

$$\tilde{l}(\beta) = \sum_{i=1}^{n} \{Y_i log\pi(x_i{}'\beta) + (1 - Y_i) log[1 - \pi(x_i{}'\beta)]\}$$

Then, the negative log-likelihood function can be expressed as:

$$l(\beta) = -\sum_{i=1}^{n} \{Y_i log\pi(x_i{}'\beta) + (1 - Y_i) log[1 - \pi(x_i{}'\beta)]\} \quad (4)$$

By using the lasso, we need to estimate β with the criterion

$$\hat{\beta} = \arg\min_{\beta} l(\beta), \text{ subject to } \sum_{j=0}^{p} |\beta_j| \leq s \quad (5)$$

Define a function

$$L(\beta, \lambda) = l(\beta) + \lambda \sum_{j=0}^{p} |\beta_j|$$

Then, the criterion (5) is equivalent to

$$\hat{\beta} = \arg\min_{\beta} L(\beta, \lambda) \tag{6}$$

where λ is a penalty parameter. By using optimization techniques, the parameter β can be estimated with the criterion (6). In our computation, β is initialized as 0. The procedure of the iterative algorithm for the activity feature selection is summarized as follow:

1) Fix λ and initialize the predictor variable set.
2) With current variables, compute β that minimizes (6), i.e. the constrained maximum likelihood estimator.
3) Compute the accuracy based on β and current predictor variables.
4) According to β, remove variables with coefficient close to 0 or significantly smaller than others.
5) Repeat steps 2, 3, and 4 until no more predictor variable can be removed.

Applying this algorithm, the significance of activity features can be measured. The most significant feature set will be selected based on the best accuracy and the least predictor variables. Even with the same predicting accuracy, the more condensed variable set is preferred. In this way, the overfitting effect could be reduced further.

V. Experiment and Validation

In this section, we validate the effectiveness of the lasso in activity feature selection on both Facebook [14]and Citeseer datasets [15]. The experimental results show that activity features can accurately predict the evolving pattern and the most significant activity features can be effectively selected by the model.

A. Dataset

We measure the evolving pattern and evaluate the lasso on both Facebook wall-posting social network and the Citeseer co-authorship network. The Facebook social network includes totally 876,993 post records between 46,952 people over 1,596 days. Each record includes two anonymous user IDs and a time stamp, indicating that the second person posts on the wall of the first person at the specified time. By setting one week as a unit and removing the records without any posting, we generate the statistics of Facebook over 219 successive weeks. The activity features and evolving patterns are extracted each week. Among all the evolving pattern labels in the dataset, there are 141 growing and 78 shrinking, which are marked as 1 and 0, respectively.

The Citeseer co-authorship network is collected from the years 1980-2006, which includes 486,324 collaboration records and 283,155 authors. Different from the Facebook dataset, the activity features and evolving patterns of Citeseer are measured annually. The maximum number of publications is 120,361 among 91,722 authors in the year 2000. In addition, the records with growing pattern are marked 1 and those with shrinking pattern are 0.

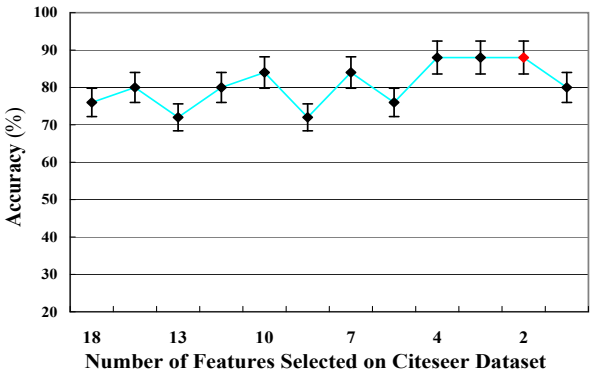

Figure 3. The predicting accuracy over different number of features on Citeseer data

B. Parameter

Only one parameter λ is involved in the model construction and feature selection. With λ changing from 0.1 to 5, we found that the fitted coefficients keep the same. Based on that, we set λ value as 1.

C. Evaluation

In the first part of the experiment, significant activity features are selected on Citeseer data. As the data only covers 25 years, relatively short for the time interval to construct ISA, the overall data is used for both training and testing. Figure 3 shows the prediction accuracy on different activity feature sets in each iteration. Error bars represent 5% errors. The best accuracy 88% is obtained when the number of features used is 2, 3, or 4. Since the set of 2 features (marked red in Figure 3) requires the least predictor variables when producing the best accuracy, it is selected as the final result.

Since the Facebook data covers a total of 223 weeks, we randomly select 53 data points for testing and use the remaining 170 data points for training. The prediction accuracy with different numbers of activity features in each iteration are shown in Figure 4. Error bars represent 5% errors. The model produces the best accuracy of 79.3% when only 2 features are included as the predictor (marked red in Figure 4). It is 5.7% higher than the accuracy obtained with full feature set. Other feature sets are not considered as good choices, because they produces lower accuracy with higher complexity. Therefore, the set of 2 features is selected as the final result.

On both datasets, although the full activity feature sets carry more information than the selected ones, they do not

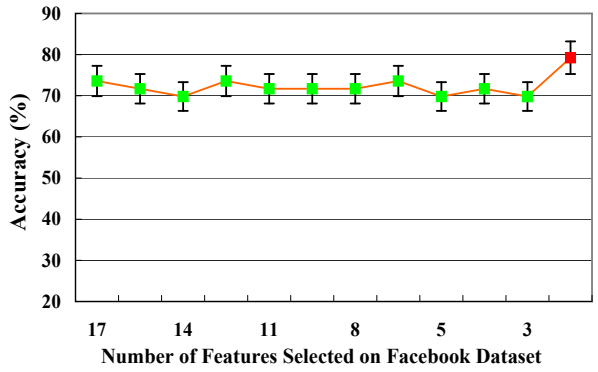

Figure 4. The predicting accuracy over different number of features on Facebook data

Table III
ACTIVITY FEATURES SELECTED AND ACCURACY ON CITESEER AND FACEBOOK DATA

Dataset	Citeseer	Facebook
Features selected	AR_t, C_t	$\|V_t\|$, CE_t
Accuracy	88%	79.3%

produce a higher accuracy. It indicates that the feature selection reduces the overfitting effects, and therefore improves the prediction accuracy.

Table III illustrates the features selected on the two datasets with their prediction accuracy. On Citeseer data, the activity features selected are the total number of current collaborations and average number of coauthors of each person. It indicates that the current collaboration between authors is the most significant factor to explain the evolving pattern of co-authorship network. On Facebook data, the features selected represent the current active population and the cumulative connections. This result reveals that in friendship network, the number of active members and their cumulative connections could be the most important factors to determine the network evolution.

VI. CONCLUSIONS

In this paper, we focus on the member activities and study their impact on the social network evolution from a macro scope. By incorporating activity features, we present a model to predict the network evolution and find out the most significant features. For future work, the evolving pattern can be further explored in different applications. Also the evolution measure can be extended with the absolute values to fit with more scenarios.

REFERENCES

[1] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "Graphscope: parameter-free mining of large time-evolving graphs," in *KDD2007*. New York, NY, USA: ACM, 2007, pp. 687–696.

[2] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM1999*, 1999, pp. 251–262.

[3] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in *WWW2008*, 2008, pp. 695–704.

[4] G. Palla, A.-L. Barabasi, and T. Vicsek, "Quantitative social group dynamics on a large scale," *Nature*, 2007.

[5] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, October 1999.

[6] E. Elmacioglu and D. Lee, "Modeling idiosyncratic properties of collaboration networks revisited," *Scientometrics*, vol. 80, no. 1, pp. 195–216, July 2009.

[7] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *KDD2008*, 2008, pp. 462–470.

[8] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," in *KDD2007*. New York, NY, USA: ACM, 2007, pp. 913–921.

[9] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *WWW2007*, 2007, pp. 835–844.

[10] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *KDD2005*, 2005, pp. 177–187.

[11] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *KDD2006*, 2006, pp. 44–54.

[12] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *IMC2007*, 2007, pp. 29–42.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996.

[14] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *WOSN2009*, 2009, pp. 37–42.

[15] C. L. Giles, K. D. Bollacker, and S. Lawrence, "Citeseer: an automatic citation indexing system," in *Prodeedings of the third ACM conference on Digital libraries*, 1998.