Studying the Evolution of Flickr Network

Yuhao Zhang
Computer Science Department
Stanford University
Email: yuhao@cs.stanford.edu

Xinxing Jiang
Computer Science Department
Stanford University
Email: jxx09@cs.stanford.edu

***Abstract*—In this paper, we propose methods and experiment with Flickr network to solve two problems. First, we study how the community in a social network evolves over time. We then report our findings about community evolution and new node behavior in the Flickr network. Second, we define the leader group of a network and propose novel mothods to measure the evolution of leader group. We then propose a machine learning-based method to predict a user's leadership status change over time. The experiment on Flickr network gives promising result.**

I. Introduction

Nowadays, social network (like Facebook, Twitter, LinkedIn, Flickr, Instagram) has been becoming an indispensable part of people's daily life. In our project, we will study how social networks and the communities on it evolve with time. Particularly, our study will be conducted on the Flickr Network. We mainly care about two aspects of the network. First, how does the network properties evolve with time? Second, how do the communites carried by the network evolve with time. Our dataset is provided by Prof. Leskovec.

The paper is organized as follows. Section 2 describes the related work. Section 3 describes the data sets under study. Section 4 shows the network property analysis in static view and dynamic view. Section 5 shows the community evolution process of the network. Section 6 shows the leader group evolution process of the network. Finally, section 7 gives conclusion and discusses the future plan of the work.

II. Related Work

Our study is a novel attempt in its area. Thus, there is no exactly-related paper talking about this topic. However, we find several papers that study the evolution of social networks and provide us with great illustrations in our research.

McAuley's paper [1] studies about users social circles categorization. This topic is connected with the topic of social circle detection that is covered in our class. The study models the similarity between one specific users friends as a function of the profile information, and proposed an unsupervised approach to learn the dimensions of profiles that lead to clustered circles.

Kossinets's paper [2] is a theoretical study on the evolution of social network - but not online social network. The author analyzes a dynamic social network comprising students, faculty, and staff at a large university. The data that are used in the study is time-stamped e-mail headers recorded over one academic year with affiliations and attributes. The key finding in this paper is that network evolution is dominated by a combination of network topology and the organizational structure of the network.

Kumar's paper [3] studies the measurements of two online social networks to learn about the structure and evolution of them. The author proposes a possible structure for online social networks, which consists of three parts - singletons, giant component and middle region. It analyzes the properties of each of these structures on real-world social networks and studies the formation and merge of these components. Finally, based on the observations, the author proposes a simple model of network evolution to capture the properties he discussed about using a small parameter space. The most important finding is that the giant component in online communities is pretty much smaller than expected - often less than half.

III. Dataset

Our dataset is extracted from the Flickr database from March 2003 to September 2005. The dataset consists of three parts: a temporal network edge list, a user registration data list and a user tagging log. The network contains 584207 nodes and 3554130 edges, with an average user degree of 6.1.

IV. Network Property Analysis

We create our graph using snap.py. The nodes correspond to users, and the edges correspond to following relations. Since following relations have direction, the graph is directed. We use registration time as node attribute and contact creation time as edge attribute. Then we analysized the properties of the network.

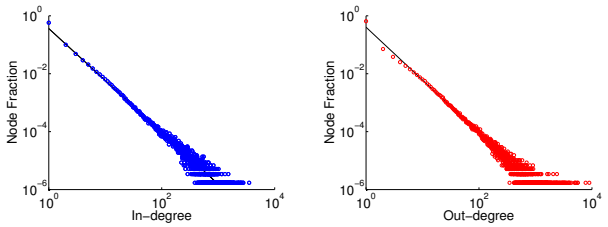
A. Static Properties

First, in order to gain an overall understanding of the network, we study some static properties of the Flickr network. The analyzed properties include basic properties like average node degree, strongly connected component size and clustering coefficient. We also study its degree distribution and node life time.

1) Basic Properties: The basic properties of the Flickr network is shown in Table I.

TABLE I: Basic properties of the Flickr network

Property	Value
Average Degree	11.1722
Strongly Connected Component Size	236340
Strongly Connected Component Fraction	0.4618
Clustering Coefficient	0.1303



(a) (b)

Fig. 1: (a) Node In-Degree Distribution, (b) Node Out-Degree Distribution

2) Degree Distribution: Since Flickr network is a directed network, we study in-degree distribution and out-distribution separately. As it is shown in the Figure 1a and Figure 1b, node degree follows power-law distribution. We use Maximum Likelihood Estimation to fit the distribution, and get α for in-degree distribution as,

$$\alpha = 1.76$$

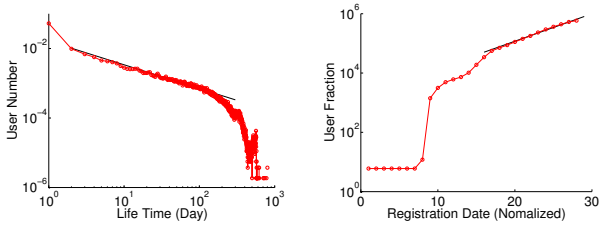
and α for out-degree distribution as,

$$\alpha = 1.81$$

3) Node Life Time: We define node life time (namely user life time) as the time between the 1st edge and the last edge of a node. By definition, it is a measurement of the users' activity. We plot the user life time distribution. It can be aboserved clearly in the figure that the user life time follows a power-law distribution with exponential cut-off.

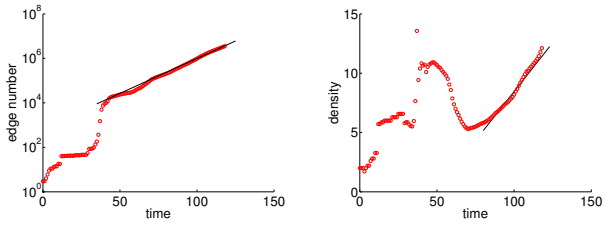
B. Dynamic Properties

We studied the evoluation of the whole graph over time. In specific, we studied the nodes, edges, density (average node degree), max strongly connnected component size, max strongly connected component fration, clustering coefficients.



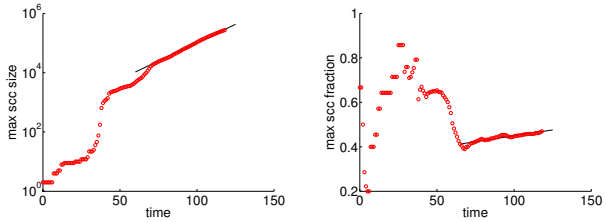
(a) (b)

Fig. 2: (a) Node Lifetime, (b) User Arriving Time



(a) (b)

Fig. 3: (a) Edge Number, (b) Average Degree



(a) (b)

Fig. 4: (a) Max Scc Size, (b) Max SCC Fraction

We have data from June 2003 to Sep. 2005. Time format of original data is second. We converted time format into week. Thus, we have about 118 weeks totally. The X axes of following analysis are all in this time format.

1) Nodes: We study the change of nodes number over time. This is the same value as the user arriving time. As it is plot in Figure 2b, the user arriving time follows an exponential distribution. Note the first several months is not stable, so the exponential distribution is clearly shown after 10 months.

2) Edges: Figure 3a shows the egde number over time in logarithmic coordinates.

Using curve fitting, we get the mathematical expression about y and x:

$$y = 719.6 * e^{0.0713x}$$

3) Average Degree: Figure 3b shows the average degree over time in linear coordinates.

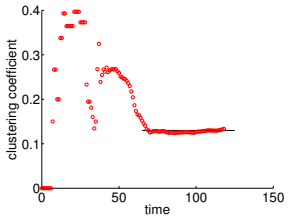
Using curve fitting, we get the mathematical expression about y and x:

$$y = 0.1647 * x - 8.013$$

4) Max SCC Size: Figure 4a shows the max scc size over time in logarithmic coordinates.

Using curve fitting, we get the mathematical expression about y and x:

$$y = 346.3 * e^{0.05696x}$$



(a)

Fig. 5: (a) Clustering Coeffient

5) Max SCC Fraction: Figure 4b shows the max scc fraction over time in linear coordinates.

Using curve fitting, we get the mathematical expression about y and x:

$$y = 346.3 * e^{0.05696x}$$

6) Clustering Coefficient: Figure 5a shows the clustering coefficient over time in linear coordinates.

As we can see, when time > 15, the relationship between clustering coefficient and time is like linear, and the slope is about 0.

V. Community Evolution

A. Community Detection

We tried to detct communities in a netwrok based on topological analysis. Then, we devided these communities into different categories according to some criterions for further studies.

B. Community Evolution Analysis

After detecting several communities, we tracked several communities over time to observe their evoluation. We selected communities with different sizes, say big communities, medium communities, and small communities. We first studied the overall property of communities. Then, we went further and studied the underlying factors behind properties.

C. New Node Behavior Analysis

When a new node N_{new} comes (i.e. when a new user registers in Flickr), it will follow some existing nodes, and join existing community. We will analysis what community a new node will join.

D. Experiment

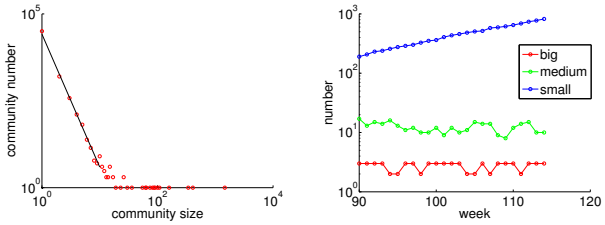
To detect communities, we tried to design an algorithm to detect communities.

However, this algorithm didn't not perform well, so in instead we implemented community detection through snap build-in function, **CommunityCNM()**, which runs fast but has a maximum process ability. So we sampled 10% nodes randomly from the whole graph and studied community evolution on the subgraph.

We tried to plot the distribution of community size at a given time. As shown in Figure 6a, the distribution of community size followed power-law.

Algorithm 1 Detecting communities in a graph $G = (V, E)$

N $\leftarrow$ the biggest in-degree node in V, $C \leftarrow \{N\}$
while Modularity(C) $>$ threshold **do**
  CC $\leftarrow$ connected component with N in G
  N' $\leftarrow$ a node in CC and not in N'
  $C \leftarrow C + N'$
end while
$G \leftarrow G - C$
repeat the WHILE loop to find next community, until find k communities



(a) (b)

Fig. 6: (a) Power-law Distribution of Community Size, (b) Community Number Evolution

In order to track different communities, we first divided communities into three categories: big community, medium community and small community. We did this according to the fraction of number of nodes in community and number of nodes in the graph. After several trials, we selected 0.04 and 0.1 to be the boundary of three categories.

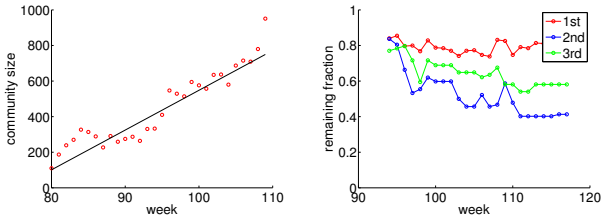
We observed the number of three categories of communities over time. As shown in Figure 6b, the number of big communities remains 2 or 3 over time. The number of medium communities fluctuates between 10 and 20 over time. The number of small communities kept increased with a constant slope over time. Since we adopted logarithmic coordinate, the number of small communities grew exponentially.

Since our subgraph is sparse, number of nodes in communities is low. After several trails, we decided to track the evolution speed of the biggest community. As shown in Figure 7a, the increase speed of node number of the biggest community is nearly linear.

We tracked node evolution of the top three communities. At time t_1, we marked all nodes in a specific community, say the biggest community, then we observed how these nodes flows at time t_2, say whether they remained in the same community or they flowed to other communities. We recorded the fraction of nodes remained in the same community and plot the curve.

As shown in Figure 7b and Table II, big communities were able to survive over time, say 23 weeks, and remained a significant fraction of initial nodes, say 40%.

We tracked the node evolution of medium communities,



(a) (b)

Fig. 7: (a) Increase Speed of the Biggest Community, (b) Size Evolution of the Biggest Communities

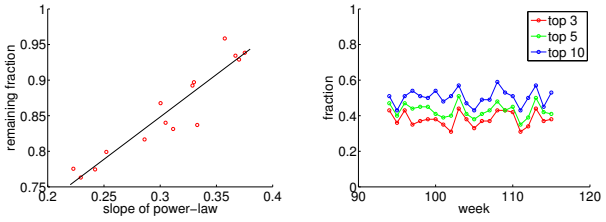
TABLE II: Size Evolution of the Biggest Communities

Initial Size	Remaining Fraction of Nodes
344	85.76%
92	41.30%
74	58.11%

and we found that they quickly extinct over time. We also tracked node evolution of small communities, surprisingly, instead of extinction as soon as medium communities, small communities survived and appeared to never changed over time, say a small community of two nodes existed from the start to the end of our experiments.

We tried to explain what factor could affect the remaining size of a community. After several trails, we found that the slope of the distribution of in-degree of nodes of a community had some effects on the remaing size of the community. As shown in 8a, there is a approximate linear relationship between slope and remaining fraction (time interval is 8 week). The possible explanation is that bigger slope means unequal in-degree distribution, which creates high rollers that help maintain a community than several nobody.

To studied new node behavior, we sampled 100 new nodes randomly at a given time, say t_1, then we tracked which community they joined at the very next timestamp, say t_1+1. As shown in Figure 8b and Table III, a significant fraction of



(a) (b)

Fig. 8: (a) Relationship Between Slope of Power-law and Remaining Fraction, (b) Fraction of New Nodes that Join Top Communities

TABLE III: Average Fraction of New Nodes that Join Top Communities

Top Index	Average Fraction
3	38.05%
5	42.95%
10	50.41%

sampled new nodes, say 40%, joined top communities at the very next time.

VI. Leader Group Evolution

Since the Flickr Network is a directed graph, and the in-degree of nodes follow a clear power-law distribution, there exist a lot of high-in-degree nodes in the graph. These nodes are significant to the properties of the network in both network theoretical study and real life scenario, so that we define them as the leaders of the network communities. In other words, we can rank the nodes by their in-degrees and define the **leadership** of a node as its rank in such a way. This leadership is significant in the way that it will influcene topics like information propagation and influence maximization, etc. Following this method, if we rank all the nodes by their leadership and take the top k nodes in the graph, we then get a **top k leader group** of the network. As the network evolves, the structure of the network will change over time. Thus, the top k leader group of the network will not stay the same. When doing this research, our question is, first, how does this top k leader group change over time? Second, what are the underlying reasons that motivate such variations.

In order to answer these two questions, we first propose our hypothesis. Then we formalize the question, propose two kind of measurement of leader group similarities and run the algorithm to compute the leader group similarities on the Flickr network to verify our hypothesis. We give many interesting results of our investigation. After that, we define a problem where we hope to predict the leadership change of every individual in the network and propose a machine learning-based method to perform the prediction.

A. Leader Group Distance

We first formalize the top K leader group that we defined in the previous part. A **top k leader group** is a group of nodes such that the leadership of every node is less than or equal to k. As it is defined in the following equation,

$$G^{(k)} = \{\text{node } n : \text{leadership}(n) \leq k\} \quad (1)$$

Following this definition, we can further define a **segmented (k,n) leader group** of a network is the group of nodes that are in the top $k \cdot (n+1)$ leader group but not in the top $k \cdot n$ leader group. In fact, segmented (k,1) leader group is just the top k leader group. Formalizing it, we get

$$G^{(k,n)} = \{\text{node } n : k(n-1) < \text{leadership}(n) \leq kn\} \quad (2)$$

In order to quantify the variation of leader group change, we propose two kinds of leader group measurement - Membering Distance and Ranking Distance.

A **Membering Distance** is defined based on the observation that during the network evolution process, there will be members moving out of or into a specific leader group, and such kind of moving-in or moving-out will change the leader group in its member composition. We define the Membering Distance between two groups $G1$ and $G2$ as

$$D_m(G_1, G_2) = \frac{|G_1 \cap G_2|}{|G_2|} \tag{3}$$

The denominator is a normalizing factor so that the membering distance between any two groups can take on values in $[0, 1]$. Note that since in most cases, G_1 and G_2 will have the same number of members, so that the membering distance between two groups has nothing to do with the order of how to compare the two groups.

Another important observation to measure the similarity between two leader groups is that, besides member composition change, a permutation of members inside the groups will differentiate the two groups dramatically. Thus we define the **Ranking Distance** between two groups $G1$ and $G2$ as

$$D_r(G_1, G_2) = \frac{\sum_{n \in G_1 \cap G_2} |R_{1,n} - R_{2,n}| + \sum_{n \in G_1, n \notin G_2} |R_{1,n} - (|G_1| + 1)|}{|G_1|^2/2} \tag{4}$$

The ranking distance between any two groups can take on values in $[0, 1]$. Note that the ranking distance is defined in such a way that a group with the inverse rank of members will have a ranking distance of exactly 1 with the original leader group; a group with all different members will also have a ranking distance of exactly 1 with the original leader group. Only a group with exactly the same members and same ranking order of these members will have a ranking distance of 0 with the original leader group. A slightly change in member composition and member ranking order will result in the ranking distance taking value in $(0, 1)$.

Both a smaller membering distance and a smaller ranking distance indicates a larger similarity between two leader groups. These two measurements are designed to be working together to show how a leader group changes over time. For example, if a leader group always have a membering distance of 0 and a ranking distance of a constant value with its past-form group, it indicates that this leader group is evolving without member composition change but with a stable rate of ranking order change. In the result part, we will show the application of these two measurements to studying the leader group evolution on the Flickr network.

TABLE IV: Features used to predict leadership change

Index	Features Description
1	Neighbor in-degree average difference at time t
3	Max follower degree change rate over t and $t+1$
4	Average follower degree change rate over t and $t+1$
5	Out-degree change rate over t and $t+1$
6	Max following degree change rate over t and $t+1$
7	Average following Degree change rate over t and $t+1$
8	Number of unclosed following triads at time t
9	Recent tag activity over t and $t+1$
10	Recent photo uploading activity over t and $t+1$

B. Prediction of Leadership Change

The study of leader group evolution has illustrated us with non-intuitive result (shown in the next part). In fact, status of the members in leader groups will change at a constant rate when the network structure evolves to a mature state. An obvious question is, what motivates users' change in their ranks? What's further, knowing some basic facts about a user, is it possible to predict whether the user is going to drop in its rank or keep at the current status after a period of time.

Considering the difficulties of making such predictions, we appeal to machine learning-based methods. And considering the fact that a user's behavior may not influence his rank shortly, we set our unit of period to one month (four weeks). In this part, we will formalize the problem, and describe the features we use to make predictions and why we choose these features. Also, we will show the classifiers we use in the study. In the next part, we will give some classification result on the Flickr network.

1) Prediction Task Definition: Suppose a user has a leadership rank (indicator of status) of R_t at time t, and a rank of R_{t+1} at time $t+1$. Then we can make prediction y about a user's leadership change as

$$y = \begin{cases} 0 & \text{if } R_{t+1} >= Rt \\ 1 & \text{if } R_{t+1} < Rt \end{cases} \tag{5}$$

Thus, a prediction of 0 means that this user has a increasing rank or stays unchanged in the coming period of time, and a prediction of 1 means that this user experiences a drop in its rank. Note that the unit of time is one month in our following experiment.

2) Feature Selection: We list the features we used to make predictions of user rank change in Table IV. Features 1-8 are structural features, which are based on network structural properties. Features 9-10 are behavioral features, which are analyzed based on user activity log. We give a brief description for each feature here.

Neighbor in-degree average difference are the average in-degree differences between this user and all the users ranked before/after this user. Thus, these are in fact two features. We set a window of a specific size, and consider only the users

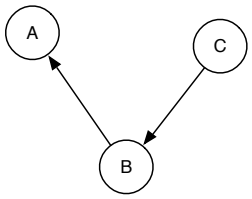


Fig. 9: An example of unclosed following triad. Node C is a potential follower of node A.

ranked before/after the user by a window size. We choose this feature based on the fact that, a user with close competitor will have a higher probability of being surpassed or surpassing others.

Out-degree change rate is simply calculated as the following equations. We choose this feature based on the fact that a user with an increasing rate of following others may get more followers over the period and earn a better rank.

$$f_5 = \frac{\text{Out-degree(t+1)} - \text{Out-degree(t)}}{\text{Out-degree(t)}} \quad (6)$$

Average follower degree change rate is calculated as out-degree change rate. We choose this feature based on the observation that a user with higher-level follower groups tends to have higher leadership, and thus ranks higher in the network leadership group. The same situation goes with max follower degree change rate.

Number of unclosed following triads is selected based on the intuition that unclosed following triads is a good indicator of future potential of being followed. A example unclosed following triad is shown in Figure 9.

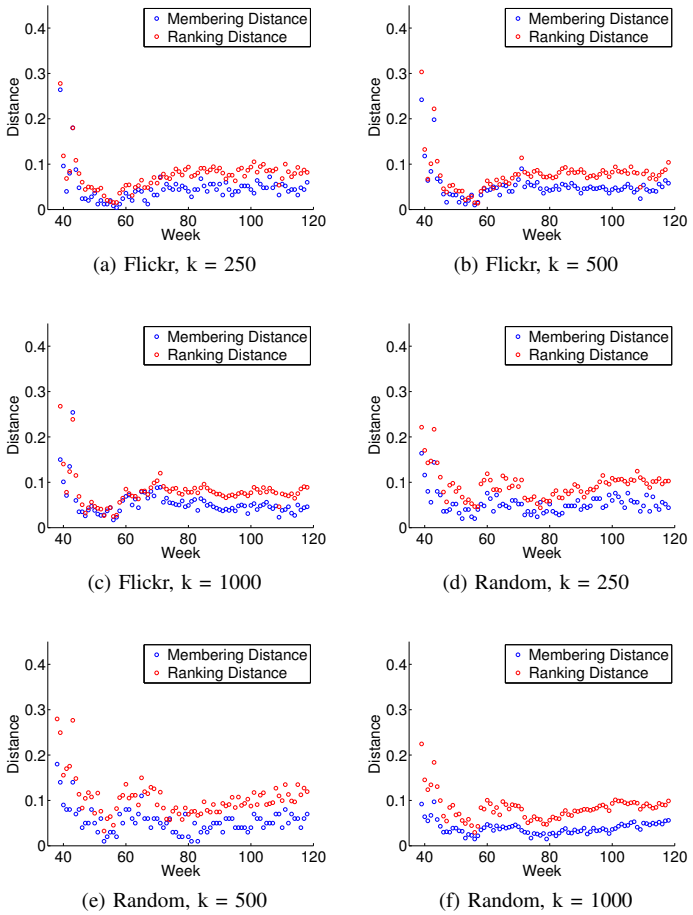
Finally, **recent tag activity** is defined as the ratio of a user's tag number over the average users' tag number in the specific time period. This is a good indicator of the users' participation and activity. The same situation goes with **recent photo uploading activity**.

3) Model Selection: We use SVM, Decision Tree and Logistic Regression as classifier models. To tune the parameters of these models, we use k-fold cross validation.

C. Experiment

After formalizing the problems we want to discover, and propose several measurements and methods, we run our methods on the Flickr network, and get many interesting results.

1) Leader Group Evolution: We have given the formalized definition of top k leader group, and defined two kind of distance between leader groups to quantify the evolution of leader groups. Now before we conduct the experiment on the Flickr network, we have several hypothesis about how the leader groups in a big social network evoles over time.



(a) Flickr, k = 250 (b) Flickr, k = 500

(c) Flickr, k = 1000 (d) Random, k = 250

(e) Random, k = 500 (f) Random, k = 1000

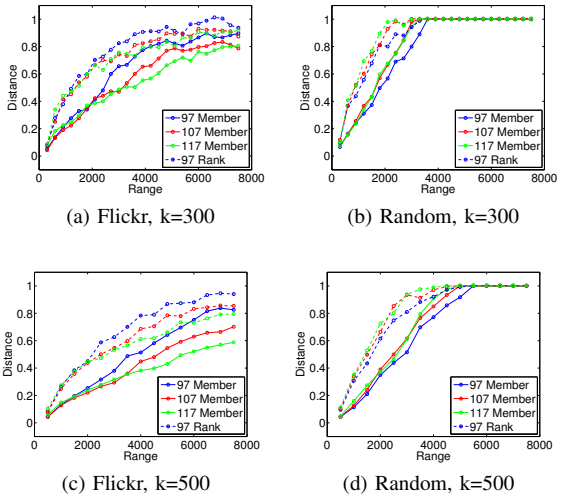
Fig. 10: caption

Hypothesis 1 is that if we fix k for a specific top k leader group, as the network evolves, the members and status (rank) of these members in this group will tend to stay unchanged. To measure this hypothesis with the formalized measurements, we expect to see that the membering distance and ranking distance for a specific top k leader group with its past form (the same leader group before a specific period of time) will tend to be zero as time goes by.

Hypothesis 2 is that if we increase n for the segmented (k,n) leader group, the members and status of members in the segmented leader group will be more and more unstable. To formalized this, it means that when we examine the distance between segmented (k,n) leader group with its past form for a larger n, then the distance will tend to become larger.

In order to verify these two hypotheses, we conduct an experiment on the Flickr network. In order to compare the result, we construct a Erdos-Renyi random network. We build the random network in such a way that the random network always has the same number of nodes and edges with the real network.

For Hypothesis 1, we run the experiment for k taking values



(a) Flickr, k=300 (b) Random, k=300

(c) Flickr, k=500 (d) Random, k=500

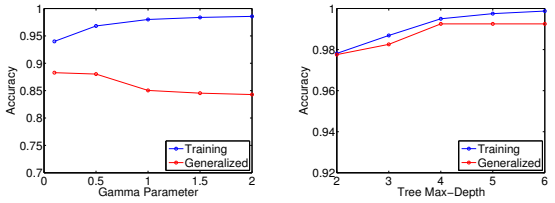
Fig. 11: caption

of 200, 500 and 1000. In order to analyze on a relatively mature social network structure, we use the networks at from 38th to 118th weeks. The results are shown in Figure 10. The membering distance and ranking distance between a leader group and its past-form group tends to be a constant value. The result is surprising in that instead of staying unchanged after a period of time, the top k leader groups tend to change at a near-constant rate. It means that the leader groups are going to reach a dynamic equilibrium in its change of members and rank of members as the network evolves. This is a non-intuitive finding. Moreover, we find that though the structure of random network shares very little in common with that of the real Flickr network, its leader group evolution is quite similar.

For Hypothesis 2, we run the experiment for n taking values of 300 and 500. In order to compare the results in different time, we use the networks at 97th, 107th and 117th weeks. The results are shown in Figure 11. The result verified our hypothesis. For the Flickr network, as n for the segmented (k,n) leader group increases, the distances grows larger and larger. This means that leaders with less leadership tend to have more unstable status. As the network evolves, this instability tends to decrease, which is indicated by the observation that the line for the network later in time (e.g. 107th week) is overal lower than the line for the network earlier in time (e.g. 97th week).

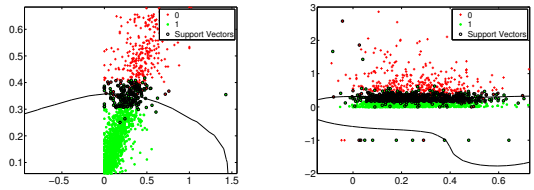
2) Prediction of Leadership Change: We experiment with the selected features and classifiers on the Flickr network. Limited by data and computation capacity, and considering the prediction feasibility, we only selected the users ranked between 100-2100 in the network for the 94th week to form our training set and test set. Thus, we have 2000 examples in total. We label these 2000 examples and randomly picks 1600 of them to form our training set, and use the other 400 examples as test set.

The classification result for SVM with Gaussian Kernel



(a) SVM (b) Decision Tree

Fig. 12: Classification result with different parameter settings. Both generalized accuracy and training accuracy are shown.



(a) (b)

Fig. 13: Decision boundary obtained by using SVM on two pair of features.

and Decision Tree are shown in Figure 12. The best result is obtained by Decision Tree.

We also show the decision boundary that is obtained by using SVM on two sets of feature pairs in Figure 13. We can observe from the figure that most of the examples can be clearly classified using SVM.

We list the best classification accuracy in Table V.

VII. Conclusion and Future Work

In this paper, we studied the structure and evolution of the popular online network, namely Flickr. Our studies analyzed the property of the network from both static and dynamic view, and discovered several significant phenomena. We detected and devided communities based on topological property, and showed the process of community evolution. In particular, we analyzed the evolution of communities of different sizes. Our work raises a number of questions about the behavioral characteristics of the new users and we will continue to study the process of new user joining a community based on user similarity.

Another main focus of our research is the evolution of the leader group in the Flickr network. We propose two novel measurement of leader group similarities and experiment with them on the Flickr network. We then report several important

TABLE V: Best Classification Result

Classifiers	Best Accuracy
SVM (Guassian Kernel)	90.5%
Decision Tree	99.2%
Logistic Regression	90.0%

findings. Furthermore, we define a new problem of predicting a user's leadership status change over time and propose a machine learning-based method to solve it. Our experiment on the Flickr network shows a best accuracy of 98% by using decision tree classifier. Limited by time and computation resources, more detailed analysis will be conducted in the future.

VIII. Acknowledgement

We express our sincere thanks to Prof. Jure Leskovec for offering us the dataset of Flickr network.

References

[1] McAuley, Julian, and Jure Leskovec. *Learning to discover social circles in ego networks*. Advances in Neural Information Processing Systems 25. 2012.

[2] Kossinets, Gueorgi, and Duncan J. Watts. *Empirical analysis of an evolving social network*. Science 311.5757 (2006): 88-90.

[3] Kumar, Ravi, Jasmine Novak, and Andrew Tomkins. *Structure and evolution of online social networks*. Link Mining: Models, Algorithms, and Applications. Springer New York, 2010. 337-357.