MEETUP GROUP LIFE CYCLE STUDY

Tongda Zhang, Haomiao Jiang

Stanford University
Department of Electrical Engineering
{tdzhang, hjiang36}@stanford.edu

*Yinan Na**

Stanford University
Department of Computer Science
nyn531@gmail.com

***Index Terms*—** Meetup, Data Mining, Life and Death Prediction, Life Cycle Analysis

1. INTRODUCTION

Meetup is online social networking portal that facilitates offline group meetings. Since 2001, there have been over 500,000 groups created on meetup, with totally over 9.3 million users registered, over 15.2 million events organized. It has witnessed a booming growth during recent years. However, among all the groups created, only part of them can stay active and vibrant for a long time while others became stagnant or dead after a while. An astonishing fact is that actually 80% of the 500,000 groups are already inactive. This pose questions to both meetup and those group owners - what cause group to die? Can we predict life and death in advance thus take early actions? Or are there any patterns of the group life cycle?

This paper is aimed to answer these questions. In this paper, we first studied about the key factors that influences the group longevity. Then we used different machining learning algorithm to predict the group life and death. Then, we analyzed the group life cycle based the group growth curve and proposed explanations to some key questions. The rest of the paper is organized as follows: In Part II, our work on group life and death prediction is presented in details. In Part III, we talk about life cycle analysis and our hypothesis. Conclusions are summarized in Part IV.

2. LIFE AND DEATH PREDICTION

Predicting the status of a group after a period of time is a crucial and important problem. Through prediction, we can not only know which group will die while others survive, but also analyze what factor is playing the most important role in predicting the result. This section will introduce the data selection and feature extraction methods and the algorithms used in prediction. After that, analysis and comments will be given according to the prediction results.

*Thanks to Professor Jure Leskovec and Professor Andreas Weigend for supervising

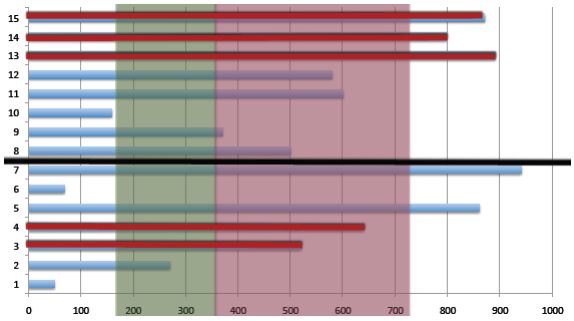

Fig. 1. Data Selection

2.1. Data selection

In prediction process, we use the data in the first year for every group and try to predict the status at the end of the second year. Before prediction, we must firstly validate the data set and choose meaningful tuples. For validation, we filtered out the tuples with invalid group id or member id. Besides, we need also check the constraints of the database, i.e. foreign key constraints. For data selection, the process is shown in Fig. 1.

As can be seen, we first align every group to the same starting point and group them into alive and dead groups. Then for alive groups, we choose to use the data of those who live longer than 2 years while for dead groups, we choose those whose longevity is between 1 year and 2 years. To make a valid baseline, we randomly choose 30,000 alive groups and 30,000 dead groups. So, our baseline is 50%.

2.2. Feature Extraction

Features to be used in prediction process are of different categories. Some features are from the original data set. These features mainly include category information of a group and location (Zip-code, longitude, latitude). Features of this features are categorical and can be used in decision tree and boosting algorithms. Some other features are the statistics of the data. This type includes group member information (member counts for each month, growth rate, derivative of

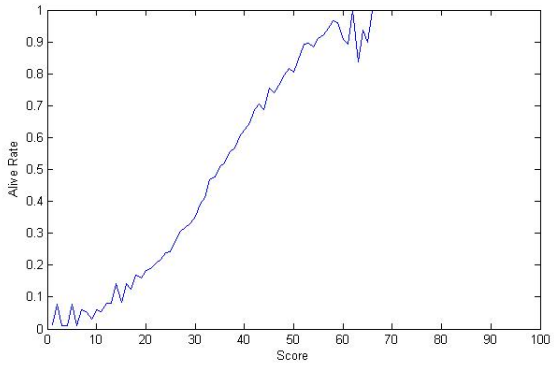

Fig. 2. relationship between the PCA score and survive rate

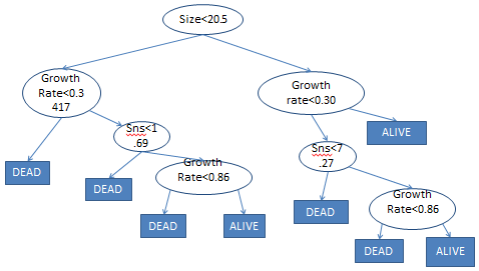

Fig. 3. Decision Tree Under 5 features

the growth rate, etc), event number, RSVP number and so on. These features are numeric and they are suitable for regression algorithm. The rest features are induced features. They are generated through experience and from mathematical computation of the data. One example is the social network (sns) feature. This feature is defined as a fraction. The nominator is the sum of times of any two members made an RSVP to a same event. The denominator is the member number of the group. This feature reveals how densely the group is connected. Besides social network feature, PCA score and spike counts are also of this type.

2.3. Algorithms

Multiple algorithms have been used to predict the life or death for the groups and we will discuss them one by one in this section. The algorithms include decision tree, linear regression, support vector machine and boosting.

2.3.1. Simple Partition

One simple idea to make prediction is simple partition. It divides the whole set by a threshold on particular variable. This algorithm can give a clear intuition and the variable dividing on can be viewed as a score to evaluate all the groups. But generally speaking, its accuracy is not high.

In our model, in order to get a better prediction accuracy, we use the first component in principal component analysis (PCA) transform of all the variables. The relationship between the PCA score and survive rate of the groups is shown in Fig. 2 . As can be seen, it's nearly a linear function, and we can say that the PCA score indicates the possibility of the group's survival.

2.3.2. Decision Tree

In decision tree, partitions are made in every node through variable from which we can make best improvement in classification. To avoid complexity and over-fitting, penalization parameters (sometimes called complex parameter) are introduced. At each node, if the improvements made by best partition exceed the threshold given by penalization parameter, we continue construction the decision tree by dividing the node. If not, we can stop and check the cross-validation / test errors. Penalization parameters are calculated by 1-std deviation criterion.

Decision tree are simple for construction/prediction. And it's easy to be visualized and understood. However, it's not very stable and the accuracy is not generally lower than the result of other algorithms. The decision tree for our problem is shown in Fig. 3

2.3.3. Regression

Regression is an approach to modelling the relationship between a scalar dependent variable and one or more explanatory variables. Due to different kernel functions, regression are categorized as linear regression and non-linear regression. In our project, we choose to use linear regression to make the prediction.

In linear regression, the prediction is made by linear predictor function. This method performs well in high dimensions and it can give us the weights of all variables. Also, linear regression can make good use of categorical variables by transforming them into multiple binary indicators.

2.3.4. Boosting

Boosting fits a linear combination on a set of weak learners to make a strong learner. The weights and parameters are decided by gradient descent. This algorithm is great in accuracy and can tell the expectation of importance of each variable.

In our model, the weak learner is the decision tree on single variable. The partition variable includes the member number, the growth rate, social network and etc. Each single predictor can achieve the accuracy around 60%.

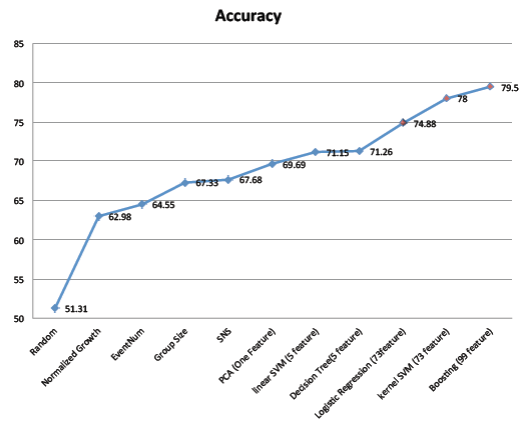

Fig. 4. Prediction accuracy for algorithms

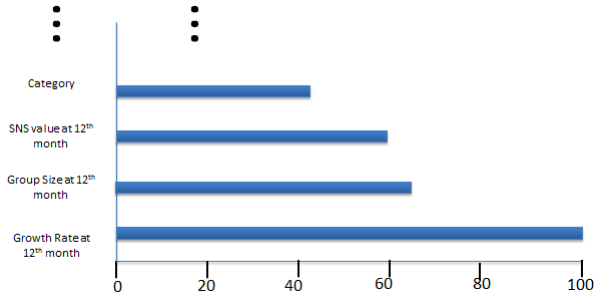

Fig. 5. Expectation Importance in Boosting

2.4. Result

The prediction accuracy for algorithms are shown in Fig. 4. As can be seen, with single predicting variable, PCA can yield the best result. The accuracy is around 71% and it's almost the same to other algorithms (decision tree, linear regression) with small number of features. With large number of features, boosting algorithm performs best and the accuracy can be pushed to around 80%.

2.5. Analysis

Besides the prediction accuracy, we can get more information about the predicting variables. From Fig. 3 and Fig. 5, we can see that the member growth rate is the most important factor to keep a group alive. If the growth rate is high, it's likely to live long, even if messages from other variables is negative (i.e. bad location). If the growth rate is slowed down, the group leader should get alarmed. Besides growth rate, group size and social network (group density) are also important. So how to attract more people to join a group and keep a good connection between group members is what the group leaders should think about.

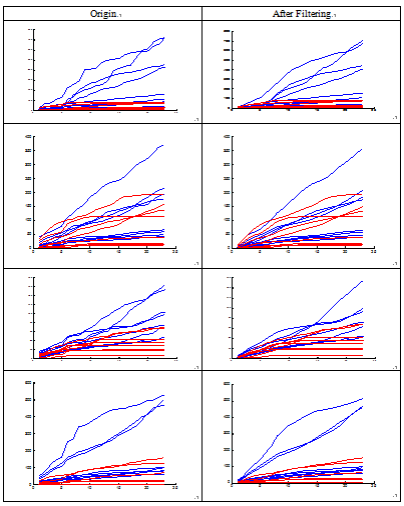

Fig. 6. Group Member Growth Pattern

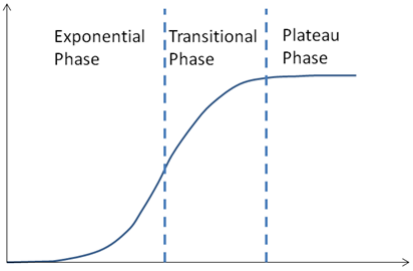

Fig. 7. Sigmoid Curve for Population

3. LIFE CYCLE ANALYSIS

After predict whether a group will die or not within a specific period, the predicting model we got shows that the group size and increase slop of a group are relatively important factors. Digging deeper about how the group size change of a groups whole life seems to be a good way to understand the group life cycle.

By drawing the group size increase curve of random selected equal number of alive and dead group, some increasing pattern can be found for most groups, no matter alive or dead. Here are some of the random selected group increasing curves. In order to prevent the random increasing noise disturb the actual pattern, each increasing curve also go through a low pass filter (here we use a 3-length average sliding window).

Seen in a big picture of the image above, especially after filtering, the group size increasing curve for both the alive groups (blue curve) and dead groups (red group) is similar to the sigmoid curve with the difference of increasing time and some other changes between each other.

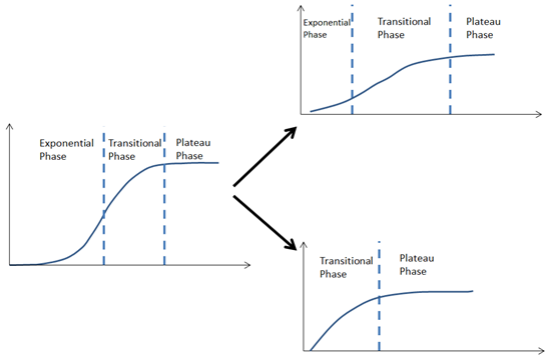

Fig. 8. Mutation of Sigmoid

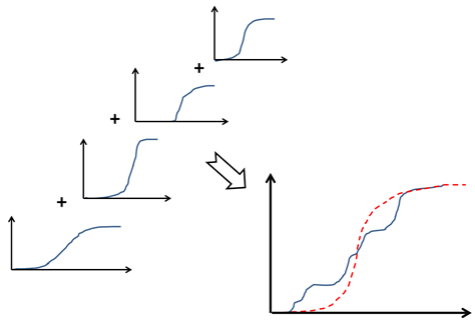

Fig. 9. Superposition of Sigmoid

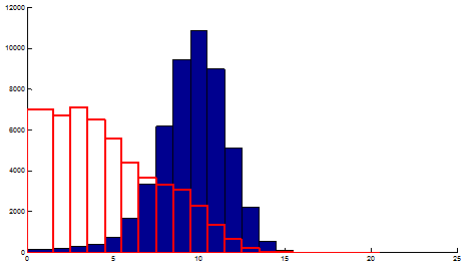

Fig. 10. Superposition of Sigmoid

Fig. 7 is a standard sigmoid curve. Knowing that the sigmoid curve, or S-shaped curve, is often used describe the population growth of a species in an ecosystem. As noted in the sketch, the sigmoid curve has three phases which are: the exponential phase, the transitional phase and the plateau phase. The exponential phase is when there is a rapid increase in population, because at the beginning of a new species coming into the ecosystem, there are abundant resources and the competition among the members are not fierce. As time passes, the transitional phase comes, that the growth rate of the species population starts to slow down. This is because the resource that no longer abundant for each member of this species. When the population exceeds a certain threshold point, the population is approaching a constant value due to the limitation of the resource and the intense competition between the members.

Compare to the standard sigmoid curve, the pattern appeared in the meet-up groups population growth curves have so many variations. Some may have very short exponential phase, some may have a really long exponential and transitional phase, and some may have very mild exponential and transitional phase (shown in belonging picture).

Another important difference is that most of the living group is consist of multi-sigmoid population growth patterns (the whole life increase curve is a big sigmoid curve).

This difference of the group size increasing pattern between living groups and dead group can be seen from the sampled groups growth curve. A statistic calculation has been done to confirm this difference. The number of sigmoid patterns within the whole life of a group is regarded as the value, then the histogram of the sigmoid pattern number for both living group and dead group is shown in Fig. 10.

The red bar refers to the dead group and the blue bar stands for the living group. This histogram clearly shows that the number of the sigmoid increasing patterns within groups whole life is very different between alive and dead group. The dead group tends to have much less group size increasing sigmoid pattern than the living group. If we set a threshold seven, the alive and dead group can be distinguished with over 75% accuracy.

Similar to the ecosystem, the reason why the meet up group increasing curve shows sigmoid pattern is simple. Consider the candidate user that may be attracted by a kind of group as the limited resource that shared by groups with similar topic, similar location, and other similar condition. At the beginning, which is the exponential phase, the group will increase really fast due to little competition. As the number of member in a group growing, less and less available candidate user will join the group, which make the increase speed slowed down to the transitional phase and then to the plateau phase. But the above explain nation only consistent with the big picture of a groups group size increasing pattern. The multi-sigmoid increasing patterns within the big increasing pattern for a group are not explained.

One possible explanation is Multi-Communities made up the whole candidate users social network that the social connection within a community is very dense, but between the communities is really sparse. When a group starts, the growth happens within a certain community, then through the sparse connection between different communities, the group growth begins in other communities, and this process is going on and on until achieve the final group size. During this process, the growth within each small community can be regarded as an independent sigmoid increasing pattern. Therefore, the whole group size increasing curve is the superposition of all the sig-

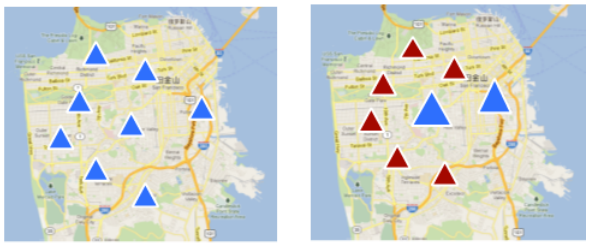

Fig. 11. 2010 vs 2012 Career/Business Meetup Group, San Francisco(blue for alive, red for dead)

moid patterns with different starting time point. This theory well explained the multi-sigmoid increasing phenomenon appeared in meet up groups. But further verification that using actually social network data to show the whole increasing social connection is needed.

Another thing we are very interested in is why only a few Meetup groups survived while most of them died. Our guess is that groups in same category competes with each other locally though they are not intended to, finally only the fittest ones survive. A typical scenario should be at very first there might be a bunch of similar groups starting in the same area. After a while, some of the Meetup group organize better events than others, thus they attracts more members. With more members, they can further organize bigger and better events and bring in more members, which forms a positive cycle. On the other hand, the other groups with less members or inferior events are actually losing their members eventually became dead.

To verify our hypothesis, we plotted the growth curves (blue for dead group, red for alive group) of all the career/business related Meetup groups in San Francisco from 2010. From the plot, we can see that most of the group stopped growing after 6 months, while several groups grows very rapidly. We further looking into the data and found that the members in these dead groups began to attend events in those alive groups after their own groups died. These facts verified our hypothesis that the locally in-category competition actually play a big role in groups lifecycle.

4. CONCLUSION

It is hard to maintain a large on-line group active for a long time. And this is especially true for the Meetup group, because it is based on the real-world group activities which is much harder to organize than pure online activities. Our work is mainly focusing on finding important indicators that decide the longevity of a group and the growth pattern of a groups development.

In part II, combining the features of group size change, social connection density, event number and other over 90 features, weve tried several machine learning method including

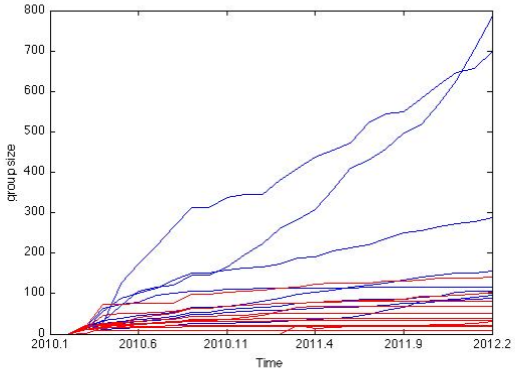

Fig. 12. Career/Business Meetup Group, San Francisco, From 2010 (blue for alive, red for dead)

linear SVM, decision tree, boosting. Our final group alive or dead prediction model can give nearly 80% accuracy by observing the group activity for 6 month, which gives us a good description of which factor or indicator plays a major role to decide the future destiny of a group.

In part III, our work starts to concentrate on the growth pattern of the group size. Our finding shows that for both living group and dead group, their growth curve is consist of several sigmoid increasing patterns, and the living groups tend to have much more sigmoid component than dead groups. To explain the phenomenon, we came up with a multi-community hypothesis, that the total group growth is consist of several small dense communities growth which give us the multi-sigmoid growth pattern. Aside from multi-community hypothesis, we also proposed a competition model to explain the phenomenon that few groups survive and most groups die in the same area for the same category. Then we verified our model by analyzing the member flow among groups.

For future research, the prediction used feature can add in social network data to further improve the prediction accuracy. And the social data can be used to verify the hypothesis we proposed in part III. Beside the social connection data, the group event data can be new target to be studied for understand the group growth pattern.

5. REFERENCE

[1] N. Ducheneaut, N. Yee, E. Nickell (2007). The life and death of online gaming communities: A look at guilds in World of Warcraft. Computer/Human Interaction 2007 Proceedings, 839848.

[2] S. Kairam, D. Wang, and J. Leskovec. The life and death of online groups: Predicting group growth and longevity. In WSDM 12, 2012.

[3] Backstrom, L., Huttenlocher, D., Kleinberg, J. (2006).

Group formation in large social networks: Membership, growth, and evolution. Proceedings of 12th International Conference on Knowledge Discovery in Data Mining (pp. 4454). New York: ACM Press.