Usage of Friendships in Social Media to Determine Lifespan

Austin Hill
a7hill@ucsd.edu
A11642378

David Situ
dasitu@ucsd.edu
A11716021

ABSTRACT:
In this paper, we will use data mining techniques in order to find structure within the Gowalla network. We will also analysis the Gowalla data to first determine if there is any pattern that can be seen. As well as compare our techniques to other techniques that are used in the industry and research..

1. INTRODUCTION

Gowalla is a location based social media site where users can share their location by "checking- in". Checking in consist of displaying to the user's friends the location with text and or other forms of media such as pictures and as well as a geolocation coordinates. Also there were "Social Guides" for individuals events or cities which were popular spots and recommendation from friends and experts for the certain location. Gowalla was a major player in the location based social media field. They were direct competitor with foursquare, another location based social media. Gowalla was later shut down on March 10, 2012 after being brought by Facebook in late 2011.

2. DATA ANALYSIS

The Gowalla's data consisted of two files with one being the friendship graph. The friendship network is directed and consists of 196,591 nodes and 950,327 edges. The second file of data set was the check-ins for every user. It contain which user check- in where and had a location ID. There is a total of 6,442,890 check-ins which was taken over the period of Feb. 2009 - Oct. 2010.

The data taken shows Gowalla as a social network emerging from infancy of a few thousand to stand at an impressive 196,000 users within a span of one and a half years. Even with a explosion of new users, average user activity over the course of the months remained relatively constant over this time. This means that most new users would have 0-9 checks in which can be seen in Table 1.

2.1 Check-In Data

Table 1: # of Check-Ins Against # of Users

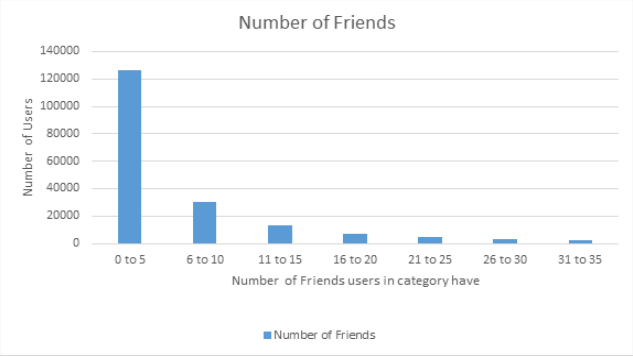
# of Check-in	0 to 9	10 to 19	20 to 29	30 to 39	40 to 49	50 to 59	60 to 69
# of User	121861	14168	18320	5215	4128	7092	2346
# of Check-in	70 to 79	80 to 89	90 to 99	100 to 109	110 to 119	120 to 129	130 to 139
# of User	4274	1425	1351	2505	890	1783	666
# of Check-in	140 to 149	150 to 159	160 to 169	170 to 179	180 to 189	190 to max	
# of User	630	1234	433	1016	347	6907	

Interestingly, the amount of users who have checked in at least once (107092 users) is barely over half of all the users in the network(196,591). This is possible because many of the users of social network are merely observers and not contributors. They want to see what their friends are doing but they personally do not add any value.

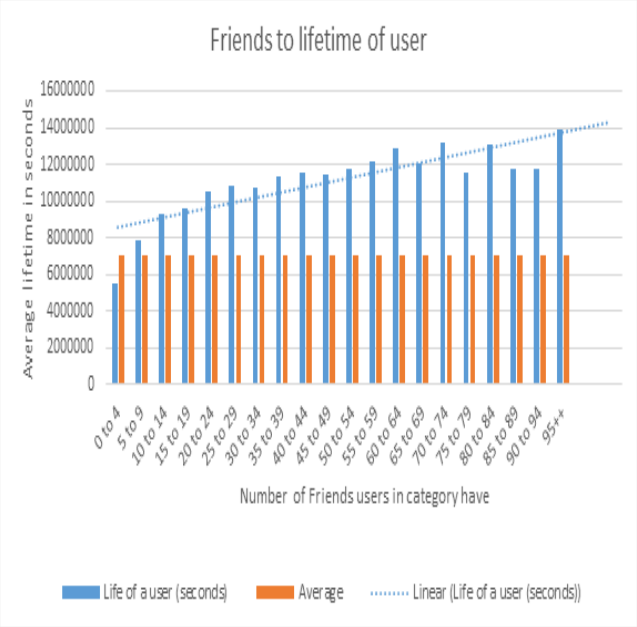
2.2 Friendship Data

Also like in most social networks, there is a bias in the distribution of the friendships in the data. As seen in Graph 1, most users have 0 to 9 friendship and then it drastically drops in the number of users that have X amount of friends. This clearly demonstrates a very sparse network graph.

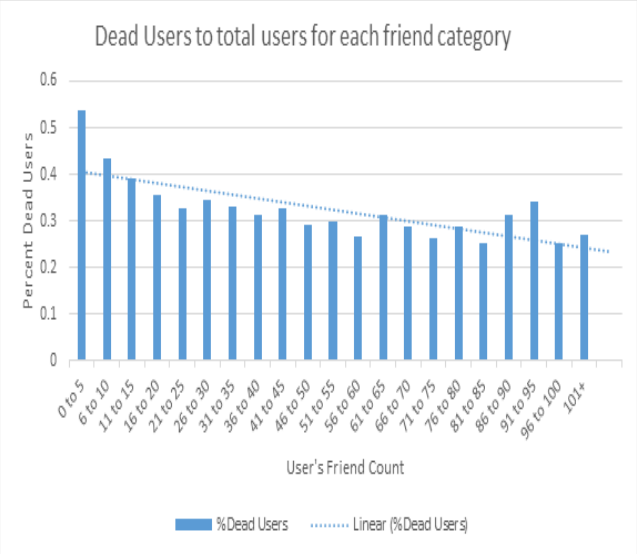
Something interesting appears when we plotted number of friendships against the lifetime of an user. In Graph 2, it shows an almost linear correlation. We define lifespan of an user based on the difference in time stamp of the last check-in and the first check – in.



Graph 1: Number of Friends



Graph 2: Friendship against Lifespan of User



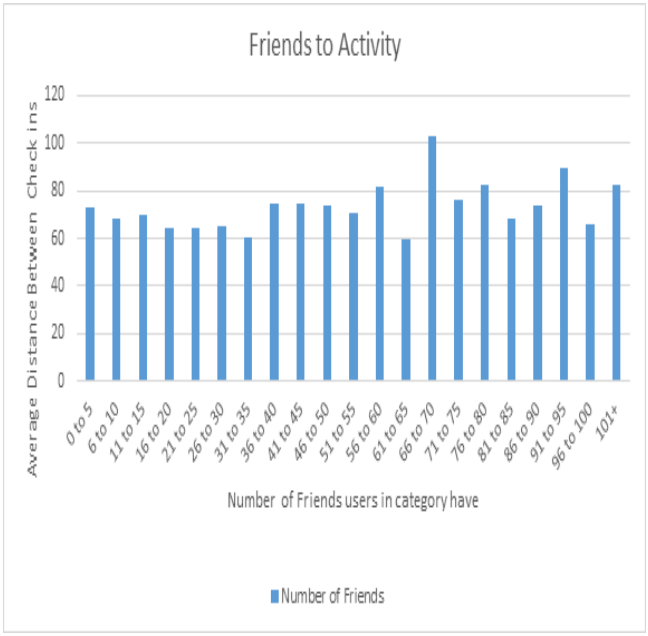
Graph 3: Friendships against Dead User

Another interesting feature of the data is the amount of "dead" users in each categories. We define "dead" as a user who does not have any check-in or their last check -in is before October 2010 because users who were using Gowalla in October are more likely to continuing past the date of data collection. As you can tell by Graph 3, users that had 0-5 friends are mostly "dead".

3. PREDICTIVE TASK

3.1 Background

What drew us to the Gowalla data set in the first place was the 1. Ease of using the data set with the check ins and friendships network, and 2. The geolocation data. We initially tried to correlate friendship to distance an user travel. Unfortunately, as Graph 4 below show there simply was not any simple direct correlation between simple distance of user movements to friendships. There seems to be no correlation between friends and average distance traveled between check ins. The reason for this might be that as users uses Gowalla more, they start to check in to more nearby places. For example, John checks in at a coffee shop then checks in at the bookstore next door. This would decrease the average distance between check ins.



Graph 4: Friendship against Average Distance Traveled

3.2 Task

After extensive (and failed) attempts to draw conclusions on the above tasks what we finally decided to attempt to model was the age of the user based on the number of friendships. How we modeled the age of the user was that if the user was active during the final month of the data collection period (October 2010) he was removed from consideration (because likely he was still an active user and his last check in wasn't really his "last" check in). We also removed users that only had a single data check in as it would be impossible to estimate their age.

This simple qualifier removed a large quantity of our data. Graph 3 shows the percent of each categories that ended in the final data set. Of the categories displayed only one kept more than 50% of its users (granted the 0-5 category also has almost 60% of the user's total). We were left with roughly 55,000 users.

3.3 The Model

We decided that the best method to model to predict how long a user would "live" based on the number of friends he had was a linear regression model. We observed that the life span jumped dramatically during the initial stage of adding friends, tapering off as the friend count increased until increasing only marginally (and sporadically unpredictably due to data point sparsity) beyond a certain point. Therefor we decided to use two different linear regressors, one for the early part of the data (1-15 friends) and one for the later part of the data (15+ friends).

General formula:

$$a_0 + a_1 * Friendship = Life\ Span$$

Where a are weights and lifespan is represented in seconds.

We took the difference in time stamp of the last check-in and the first check–in to calculate the lifespan of each individual user.

In order to measure error in our model we used a simple MSE rating (the difference between the models predicted life span of the user in seconds and the user's actual life span in seconds) compared to the baseline model of simply comparing each point to the average life expectancy. We split the data by putting users into two groups 90% training and 10% test.

4. LITERATURE

4.1 Related works:

In the paper[1], *Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities* Jiang Yang , Xiao Wei , Mark S. Ackerman, Lada A. Adamic(2010), they performed analysis on the lifespan of the user in online knowledge sharing communities. They used Yahoo! Answers, Naver Knowledge iN, and Baidu Knows as their data set. An online knowledge sharing community is a site where users post questions and other users reply with an answer. They used survival analysis to measure if the user will "survive". Survival means that the user will continuing participating on the site by either answering or asking questions. They define the start of the user's life as the user's first post. They define "death" of the user when the user is inactive for more than 100 days. They also split the data into two parts. One being the first year the site launched and the second year. The reasoning behind this split was that idea that in the second year, there is more content and users so user's survival rates will change.

They used two different variables to model the user's survival rate. One is the ask/reply ratio(A/R ratio)

$$\frac{\#\,Question\ Asked}{\#\,Question\ Asked + \#\,Question\ Answered}$$

and net points which was a measure of points gained from from answering and asking questions.

4.2 Results

After running regression on the data, they came to the conclusion that A/R ratio is a good predictor of user lifespan. The higher A/R ratio correlated with a shorter lifespan. This would logically make sense because people who ask questions are looking for a quick answer and did not contribute anything to the community compared to the user who answered. Also most user would post only one questions and that would be the only interaction from the user.

The way net points variable was generated varied across platform. Yahoo! Answer gave points for answering questions but took away points when asking questions. Nave Knowledge and Baidu Knows both gave points for both answering and asking questions. This create an interesting phenomenon where Yahoo! Answer had more users that produced

more answers compared to Nave Knowledge and Baidu Knows.

The study's conclusion are similar to our own study. Even though the variables are different, the data showed similarities. One being the phenomenon of users only using the platform once. Most the study's users used the online knowledge sharing communities for only one question. Exactly the same as our users where most would check in less than twice.

5. RESULTS

In our experiments, we created a baseline classifier to compare our own classifier against. Our baseline was a very simple model which took the average of the lifetimes of users.

5.1 Linear Regression Model

In our experiment where we tried to relate number of friends and lifetime of users in seconds, the resultant equations for the linear regression:

For number of friends less than 15:

$$Life\ Span = 4.54 * 10^6 + 4.62 * 10^5 * [\#of\ \ friends]$$

For number of friends 15 or more:

$$Life\ Span = 1.08 * 10^7 + 1.54 * 10^3 * [\#of\ \ friends]$$

This results matches what is expected from the graphs.(Graph 2) For users with less than 15 friends, the amount of time spent using Gowalla increases rapidly for every additional friend. The other interesting part is the lower starting point because most users in the lower range of friends are less likely to continue using Gowalla.

5.2 Error

To measure the error of the regressor, we used the MSE.

$$MSE = \frac{1}{N} * \sum (Y - X)^2$$

where N is the amount of data, Y is correct value, and X is estimated value.

The resultant MSE is $4.11*10^{13}$ and the baseline MSE is $4.51*10^{13}$. This showed that our linear regressor is slightly better than our baseline by a 10% improvement of the MSE. Though the reason why the MSE of our regressor is only slightly better is due to how the data is shaped. As talked about in previous sections, users who had more friends were more likely to have been using Gowalla after the last recorded data point. This causes a lot of users to be not counted as dead.. Those users would have boosted the average lifespan of the category. Another reason is that most of the dead users were in the category 0-4 which means that rest of the categories had few data points to train on. This would have made the few popular users hard to train the regressor for.

6. CONCLUSION

Through our study of the Gowalla network, we have seen interesting trends and some uninteresting trends as well. We can conclude that our model of using number of friends to estimate lifespan is only slight better than averaged lifespan baseline. This was caused by the amount of users that were not dead. This caused some bias in the data which reflects in the data graph's non linearity. In order to truly see if our model is reasonable, we would require more data points.

7. REFERENCE

[1] Jiang Yang , Xiao Wei , Mark S. Ackerman, Lada A. Adamic 2010. Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1466%40article/1856.