# Final Project - Social and Information Network Analysis

## Factors and Variables Affecting Social Media Reviews

Humberto Moreira
Rajesh Balwani
Subramanyan V Dronamraju
Dec 11, 2011

## I. Introduction

### Problem Statement

The “wisdom of the crowds” is leveraged within online communities both as a means of generating content as well as a means of evaluating it.

In this context, we intend to study issues associated with the nature and influence of rating and reputation systems in online communities and social networks, basing our work on a dataset from StackOverflow, a technology-oriented Q&A site.

Our main pursuit is to test is the following: Do reputation systems (structure/points/levels) create an entrenched set of users whose rating is subject to preferential attachment, resulting in a few influencers with a disproportionate impact on the overall network?

**Primary Issue: Does seniority (in terms of elapsed time as well as accumulated participation) create a participation bias?**

Our intuition is that there is increased connectedness among contributors with legacy status. We believe this analysis can provide insight on how influencers are formed on social Q&A and content evaluation sites.

This kind of analysis can also allow us to explore what the effect of social influence is on ratings on social Q&A sites. We will also explore added social influence of top influencers (defined by degree or status).

**Secondary Issue: Does the Pareto Principle (the 80-20 rule) hold true?**
In case of the reputation system based sites, our intuition is that the Pareto Principle holds true as a reflection of preferential attachment. The top 20% of the influencers could well exert influence over 80% of the network. A top influencer is typically an key part of a super connected cluster. A union of several super connected clusters will cover a substantial majority of the network. An analysis of the union of several influencers and the extent of their influence will reveal the applicability of the Pareto Principle.

This kind of analysis allows us to explore the effect of individuals and a potential collection of individual who could influencer a substantial majority of the network. Will this change the way a

social reputation network is designed, or will we seek ways to spread or further concentrate the influencer of top influencers? These are interesting future research questions to explore.

## II. Review of the relevant prior work

In considering prior work, we chose to look at both work that related specifically to online communities and knowledge sharing as well as literature related to graph and network theory.

On the community side, one prior work related to Q&A based content sites is the paper *Knowledge Sharing and Yahoo Answers*. The authors examine the diversity of questions being asked, breadth and quality of the answers and then predict which answers are most likely to be rated the best. The authors classify the user response from most active categories using k-means clustering based on the thread length (number of responses), content length (length of response) and asker/replier overlap. The user response data is used to create askers and repliers as graph nodes, in order to study whether communities are being formed.

Another paper is on Collaborative filtering by Koren [2]. In this paper author uses Collaborating Filtering (CF), where past transactions are analyzed in order to establish connections between users and products. Our intent was to use history from StackOverflow to analyze and establish connection between users and responses/evaluations.

On the network theory side, Erdos and Renyi's classic Random Graph model [3] provides a baseline, which we attempted to adapt to the StackOverflow data as a control model. The authors propose models for creating graphs based on collections of potential graphs with characteristics in terms of number of nodes and number of edges, as well as a model for probabilistic construction of graphs. We will employ variants of these techniques in our analysis.

As for a network model that is consistent with preferential attachment, A.L. Barabazi's [4] work finds a decaying power law distribution for for large networks, which we see as a promising way to analyze the StackOverflow dataset. Barabazi's assertion that the random graph theory may not accurately represent real networks turns out to be borne out by our social community analysis.

One extension of Barabazi's work is the 'Positive-Feedback' preference model proposed by Zhou [5] which extends the model to suggest that a stronger preference for high-degree nodes that can be modeled with an additional parameter.

## III. Network Characteristics

To gain more insight into the structure of the Stackoverflow network we have chosen a set of graph properties which represent especially telling aspects of the network. We looked at following characteristics for each of the above 3 relationships separately and combined together.

- Degree distribution
- Average degree

- Maximum degree
- Degree with k = 1, 2, 3
- Age of the node (user creation date) vs. user reputation
- Age of the node vs. degree (All types)

## IV. Actual Network

We observed three varieties of relationship on the Stackoverflow network:

1. User A creating a question and requesting a response and user B responding to that question [Type 1].
2. User A creating a question and requesting a response and user B providing a response which was selected as correct response [Type 2].
3. User B selecting question created by user A as favorite [Type 3].

Based on these characteristics, we modeled the set of interactions between users into a network graph that incorporates each of these interactions as edges between nodes.

| | All Types | Type 1 | Type 2 | Type3 |
|---|---|---|---|---|
| Number Of Nodes | 370,214 | 161,558 | 343,969 | 124,662 |
| Number Of Edges | 3,686,038 | 809,454 | 2,265,927 | 774,497 |
| Average Degree | 19.87 | 10.02 | 13.17 | 12.31 |
| Maximum Degree | 8733 | 4900 | 4723 | 3896 |
| Degree k = 1 | 1.71% | 3.73% | 2.82% | 2.87% |
| Degree k = 2 | 0.65% | 1.39% | 1.07% | 1.2% |
| Degree k = 3 | 0.43% | 0.81% | 0.67% | 0.68% |

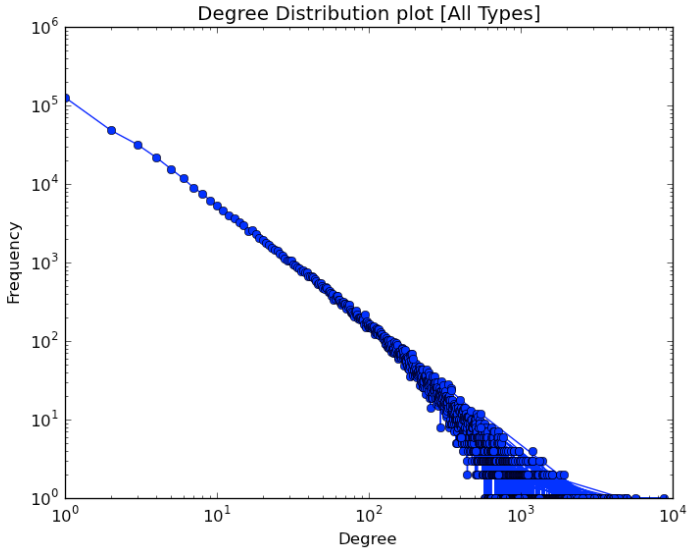
Additional Attributes

We also took additional attributes from the Stackoverflow data in order to determine the impact they have on the vote and edge data: Age, Access Date, Last AccessDate, Post Count, Comment Count, Vote Count. Upon deeper study we may find some of these attributes to be of greater significance than others.

We will describe some basic characteristics of the graph representation of the real StackOverflow data before proceeding to use this information to consider potential behavior in the community as well as generate a synthetic model that contains similar characteristics.

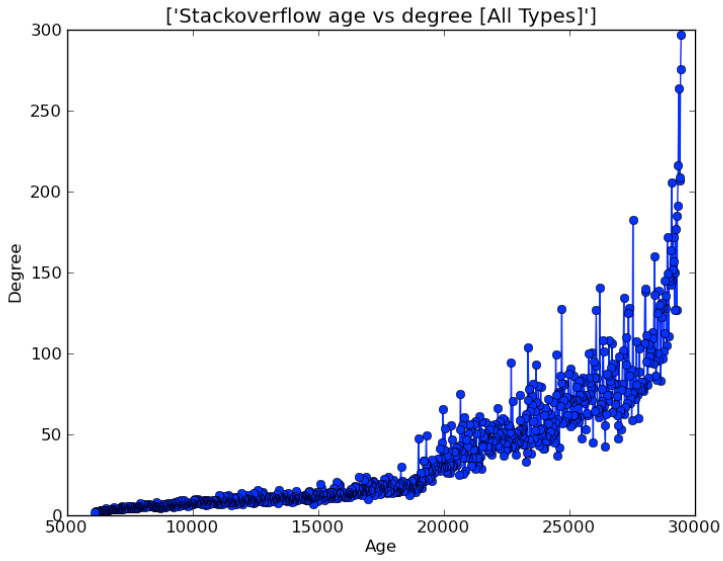
The graph representation of the StackOverflow network has the following characteristics:

**Figure 1: Degree Distribution [All Types]**



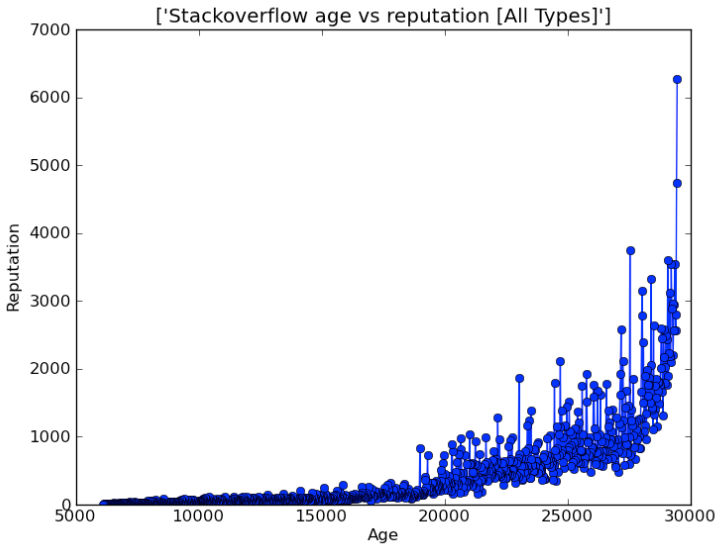

This initial look at degree distribution suggests power law behavior [Figure 1].

**Figure 2: Node Age vs. Degree [All Types]**



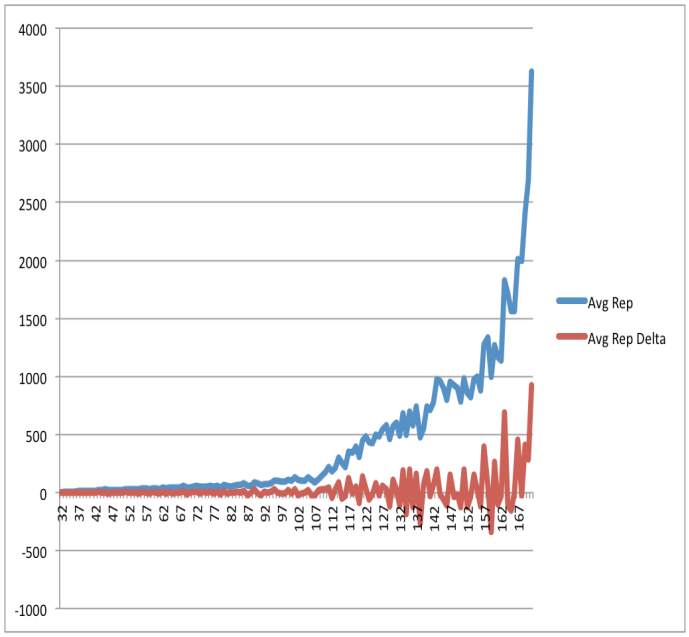

The effect of age on degree is also evident from this comparison.

**Figure 3: Node Age vs. Reputation [All Types]**




The effect of age on “reputation” a quantified metric that consists of a sum of positive feedback on the site, is also visibly related to age.

**Figure 4: Average Reputation Per Week**




One additional view allows us to examine two factors - the average reputation per week and the change in average reputation from week to week [Figure 4]. Two things are significant from this comparison, one is the fact that the average reputation does significant increase on

average for those in the oldest groups, but also the rate of increase consistently increases, suggesting the possibility of preferential attachment.

## V. Theoretical Foundation for Modeling

**Fitting the observed data from Stackoverflow.**
In order to understand and have some quantifiable method of comparing how well a model fits the underlying data; we used Least Square and Maximum likelihood to fit the observed data from Stackoverflow user graph. Using Maximum likelihood we found alpha to be **3.14** and xmin to be **227**. Whereas using Least Square we found alpha to be 0.44 with intercept at 393 and error rate of 0.21.

In order to create a synthetic model, we used 2 approaches:

1. **Barabasi Albert Model**

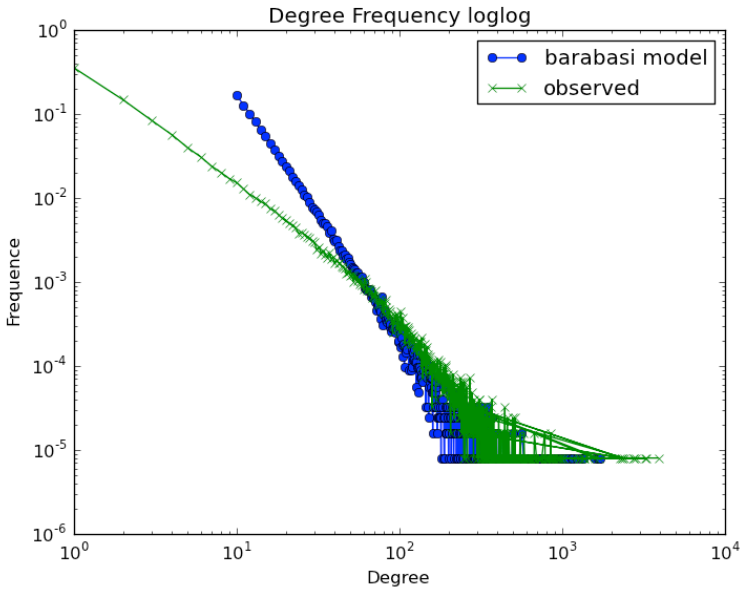
Since we have seen that our observed network from Stackoverflow user graph exhibits scale-free properties [Figure 1], we tried to use some scale-free network generation algorithms to try to model it. One of the first models that we attempted to model an emerging power law effect was the Barabasi Albert model of preferential attachment. For any given user that is asking question on the Stackoverflow site, it seems reasonable that those nodes with a high degree count (i.e. those nodes which are already well-connected) should be more likely to his question.

Barabasi model is generated by starting with an initial core of m nodes, and at each step creating a new node with m edges to already existing nodes. The result is an almost certain creation of a few highly-connected nodes (reflecting the phenomenon of few celebrities being highly-connected in social networks), and results in a degree distribution that is more like the one observed in social networks or the Internet in that it obeys the power law.

Formally, the probability pi that the new node is connected to node i is

$$p_i = \frac{k_i}{\sum_j k_j},$$

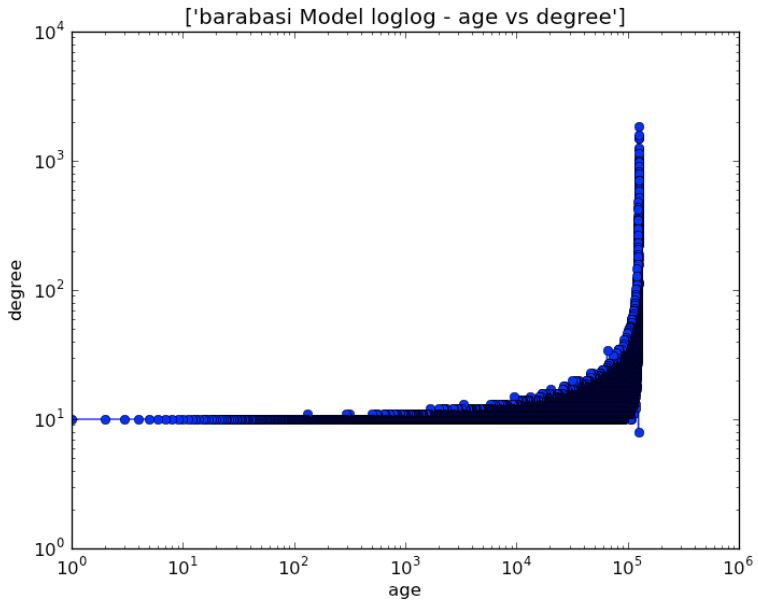
**Figure 5: Degree Distribution Observed Data vs. Albert-Barabasi Model**




**Simulation Results from the Model**

- This model strongly suggests that the graph generated from Stackoverflow favorite follows preferential attachment [Figure 2].
- The degree distribution follows similar frequency as the observed Stackoverflow data. Displaying power law characteristics, including a long tail [Figure 5].
- We also observed that we didn't see a significant number of people with very low degree when compared to Stackoverflow data. This might be due to that fact that there are many users who signed up to ask questions and there might also be a number of users who have abandoned the site over time, whereas in Barabasi's pure model, all nodes are still active.
- We saw top 5% of the users have around 85% of the overall reputation given on Stackoverflow site. Thus forming a cluster at the tail of the distribution. We observed similar cluster with Barabasi Albert Model.
- One of the problem observed with Barabasi Model was for values of m > 1, the resulting models do not have any leaves, since each created node already has several edges. For m = 1, the resulting model was too sparse and resulted in high-degree nodes that did not have a sufficient number of edges.

**Figure 6: Degree vs. Age Barabasi -Albert Model**

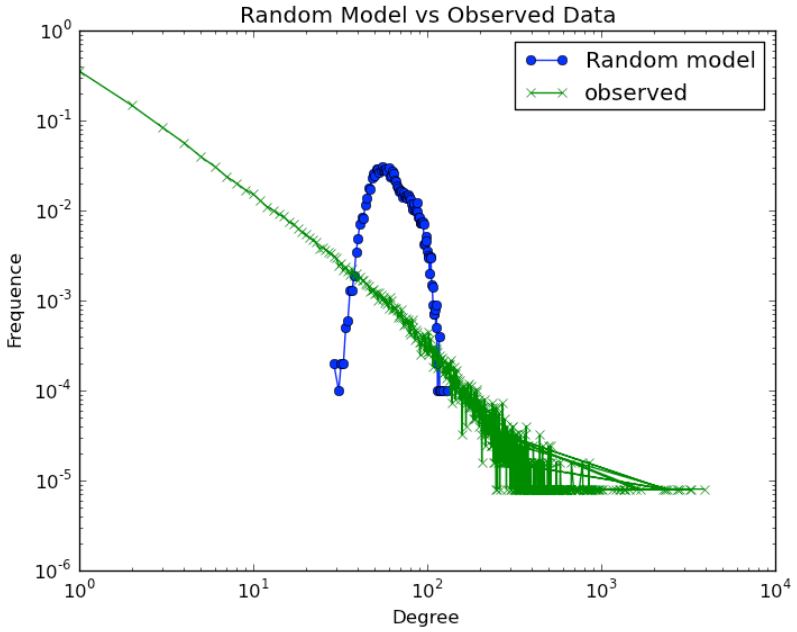



## 2. Random Model

The second method we employed to create a scale-free network for our model involved attempting to introduce randomness into the process of including new nodes.

Approach 1: Introduce one node and connect to one existing edge at random.
There was no long tail when compared with observed stackoverflow degree distribution data. Also there were a lot of users with low number of edges and less connectivity.

Approach 2: Introduced one node and connect to one existing edge at random, then connect the new node to add existing edges with some probability drawn from some uniform distribution. This model includes parameter for number of nodes to be created and the probability. This model resulted into a normal degree distribution. This was clearly different then the observed data.



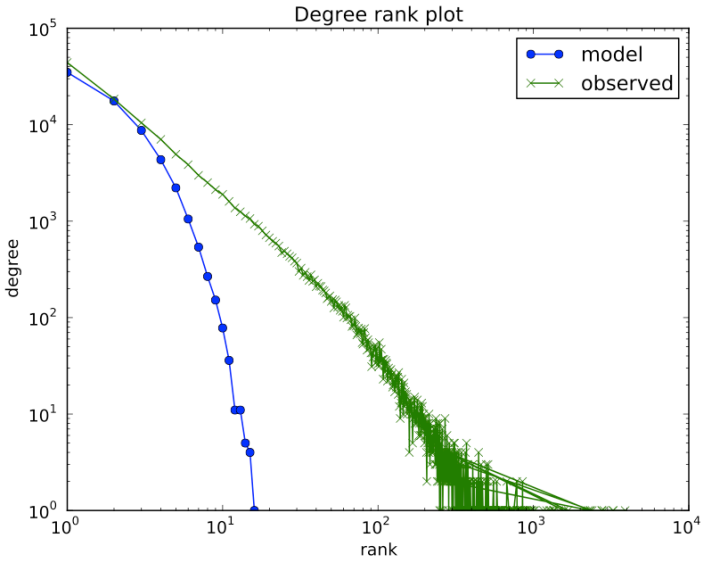

Approach 3: Introduce one node and connect to one existing edge at random, in addition connect all exiting nodes to each other with some probability.

This mode included a parameter for the number of nodes to be created and a probability with which existing nodes will be connect. We observed that this model also resulted in a normal degree distribution

Approach 4: Finally we included a parameter for the number of edges that would be connected between the new node and existing nodes, in order to tune that parameter and approach the StackoverFlow Data.

We reduced the value for number of edges to connected to as more nodes were introduced, trying to reproduce behavior of the kind where new users would be less connected than older ones. We observed that degree distribution was more clustered. We didn't see any few nodes with very high degree frequency, which was the case in the StackOverflow real data. The algorithm resulted in a minimum number of edges, but no long-tail behavior.

**Figure 7: Degree Distribution Observed Data vs Random Model**




# VI. Summary of Results and Findings

## 1. Findings from the Model

In summary, our experiments with applying both random and preferential models to the StackOverflow data showed that the preferential models clearly had a better fit to the real data, both in terms of overall power-law distribution of degrees as well as to the effect of age as a parameter. Age is clearly an important factor in determining how likely a user will be have accumulated interactions on the social network.

We additionally saw evidence that it is possible to apply the PFP (Positive-Feedback Preference Model) in order to further better fit the Barabasi model to the observed data from Stackoverflow. The BA model we choose is a more linear model where new node is connected to m existing nodes, but through using PFP we should see the disproportionate connection to existing high degree nodes and high frequency leaf nodes.

## 2. Exploratory findings from top users on Stackoverflow.

One of the hypotheses we made earlier on was whether Pareto Principle (the 80-20 rule) held true for Stackoverflow's user graph. We looked at the user reputation of all the StackOverflow users (total reputation 81,686,853) and discovered that

- Top 10% of users have 93% of total reputation
- Top 5% of users have 85% of total reputation
- Top 2% of users have 71% of total reputation
- Top 1% of users have 58.5% of total reputation

This shows a skewed distribution of ratings on Stackoverflow. The April data dump we got had 559,803 users and out of which around 55,000 users have 93% of the total reputation. This indicates that Pareto Principle does hold on Stackoverflow.

Another hypothesis that we wanted to look into was whether seniority (in terms of elapsed time as well as accumulated participation) creates a participation bias. We discovered that top 1% of users (top 1% of the site reputation) on average have an age of 974 days (23377 hours from create date) on average top 5% of users has age of 902 days (21671 hours from create date), top 10% of users on average has age of 843 days (20235 hours) and average of all the users was 552 days (13261). This indicated the users, which have highest reputation, were on average the oldest users.

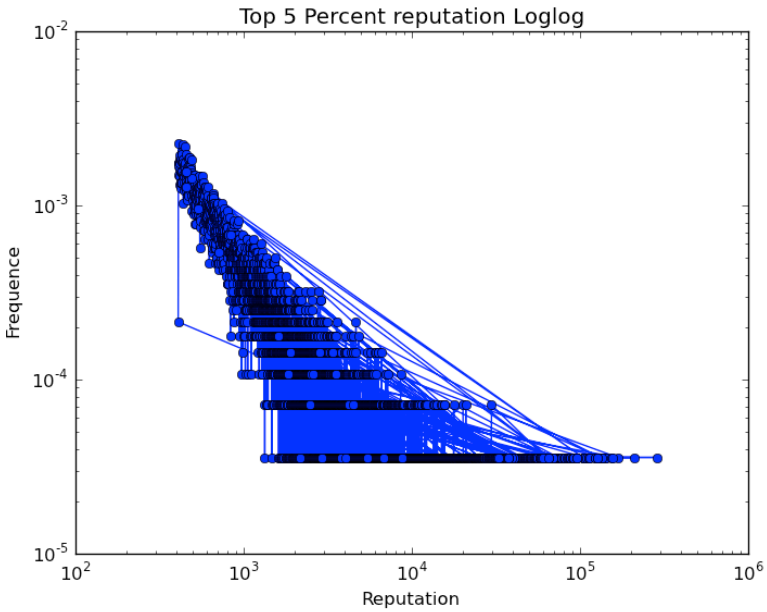
We also observed that top 5% of users attained most of the reputation from few tags. That is even though these users had on average more than thousand tags but most of the reputation was related to the top 5 tags. Which might indicate that the users were concentrated on certain areas while answering question and also exhibits user expert groups being formed.

### 3. Future research areas

One potential area for future research might be to look at why even when ratings can be both positive and negative, ratings tend to be overwhelmingly positive, unlike what might be expected from a market mechanism where different opinions create a convergence on a certain rating.

| | Up Votes | Down Votes |
|---|---|---|
| Top 5% users | 92.8% | 7.1% |
| Bottom 95% users | 97.55% | 6.3% |
| All Users | 93.6% | 6.3% |

**Figure 8: Reputation of Top 5% of Users**

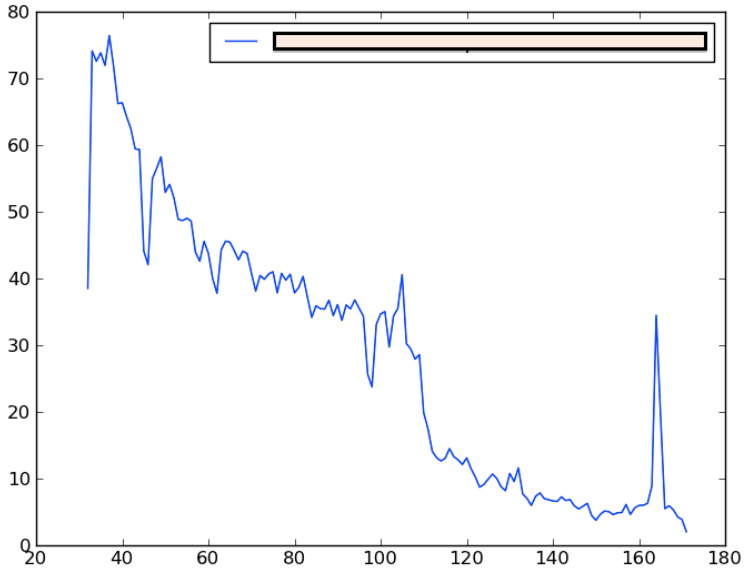



A potential future research area could be to focus on the absolute top users, the top 5%, and study how their behavior differs from the general community population.

We also explored Positive-Feedback Preference model (PFP) to avoid the rigid structure of the simulated Barabasi network. Zhou's PFP model provides an alpha parameter which increases the preference for nodes with greater degrees, so we expected that this might replicate StackOverflow's bias towards nodes of greater age.

We attempted to set the value of alpha to 0.48 to calculate the probability for new node connection to existing nodes. However, we observed that the model generated was not similar to observed model. The generated model exhibited behavior where few nodes were highly connected and extremely high number of leaves when we were connecting new node to one existing node. We also had some difficulty in sampling from a distribution that approximated that suggested by Zhou's probability density function [5]. However, we believe that there is promise in this model and it can be further explored to connect new node to more existing nodes and also to connect existing nodes with different probability than new nodes.

**Figure 9: Number of User vs. Age [Weeks]**




In the case of the source data, comparing the distribution of users' age, examining the number of users (y axis) who are a certain number of weeks old (x axis) shows a large number of users arrived in week T-165, followed by several weeks of low growth, and then a relatively stable level of growth from week T-100 onward to the present day. Further analysis could look at the cause and implications of these growth spurts in online communities.

## VII. References

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In Proc. 17th International World Wide Web Conference, pages 665–674, 2008.

[2] Y. Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. TKDD, 4(1), 2010.

[3]Erdős, Paul; A. Rényi (1960). "On the evolution of random graphs". Publications of the Mathematical Institute of the Hungarian Academy of Sciences

[4] A. L. Barabazi. R. Albert. Emergence of Scaling in Random Networks. Department of Physics, University of Notre-Dame. 1999.

[5] S. Zhou. R. Mondragon. Accurately modeling the Internet topology. Physical Review E, vol. 70, no. 066108, Dec. 2004

[5] Jure Leskovec, Jon Kleinberg, Daniel Huttenlocher, Ashton Anderson. Effects of User Similarity in Social Media, 2011

[6]Stackoverflow “About Stackoverflow”. Accessed April 2011 http://stackoverflow.com/about