# CS224 – Final Project Report
# A Study on Social Behaviors in Online Communities

Chi Wai Lau (chiwail), Xin Ai (xinai)

**ABSTRACT**
In this report we discuss the reputable users in online communities and their behaviors from different perspectives. We investigate how to label reputable users with the online activities of the whole community and compare the result to local communities and different centrality rankings. We discuss the writing behavior of reputable users, how they differ from other users and how they change in writing over time. We also build a predictive model to detect potential reputable users when they just join the community, and a generative model of the network for simulation studies.

**INTRODUCTION**
User-contributed content such as posts, comments and votes are essential to the success of any online communities. Prior work has been done on analyzing antisocial behaviors and detecting if a new member would grow into a troll based on their postings. In the paper, we would like to extend the study of antisocial behavior in online communities and perform in-depth studies on the reputable users. First we want to understand their behaviors and use network-based ranking algorithms to identify them for answering questions like: Do they also write differently over time? How do top users gain their reputations? For both reputable users and trolls, we also want to understand how they affect the overall community growth. Finally we use generative models for simulating different communities and evaluating our algorithms.

**RELATED WORK**
Prior work has been done on identifying users with expertise automatically using network-based ranking algorithms such as PageRank and HITS [3]. The network's structural characteristics are proven to be valuable for identifying the expertise level of a user. The evaluated algorithms performs almost as well as the human raters. There are pros and cons for different algorithms. Sometimes the simple one could be as good as the more complex ones. Based on simulations, the performance of the evaluated algorithms can vary when the structural characteristics of the online communities are different; therefore, we should first understand the network structure and then determine which algorithm to use for optimally labeling the user expertise level.

A big obstacle in data science is obtaining quality real world data. Simulation is very useful for understanding expertise networks and designing algorithms. Without requiring interventions in real organizations and unobtainable experimental conditions, the simulations helped us develop a good understanding of how different algorithms behave. Therefore, we need a generative model for the simulations that can mimic the growth of an online community. We can either try to

find an existing model that fits the real data, or if we do not find anything, we can build our own model and use it in the simulations.

Recent work has been done on understanding antisocial behavior in online discussion communities. Internet trolls were examined over time to understand how they eventually resulted in being banned. In summary, here are the highlighted characteristics of trolls:

- While all users write worse over time, the change in quality is larger for trolls however. Trolls also tend to make less of an effort to integrate or stay on-topic. They don't necessarily use more negative words, but they definitely uses fewer positive ones.
- The way trolls generate activity around themselves depends on the community. They are more likely to reply a post in some communities, but in others, they tend to start new discussions.
- Community tolerance changes over time. Posts written later by trolls are more likely to be deleted, regardless of whether they are actually worse.
- Excessive censorship causes users to write worse.

Based on the studies above, an algorithm was designed to predict whether a member of an online community would be a troll using only 5 - 10 online posts.

To truly understand how an online community operates and grows, we believe that understanding the good users is as important as the bad ones. We propose a similar study on the reputable users in an online discussion community. First we will use the network-based algorithms such as PageRank and HITS to identify those users. We then study their behaviors over time and design algorithms that would help us detect if a new member would grow into a reputable user. Finally we use generative models for simulating different communities and evaluating our algorithms.

**DATASET DESCRIPTION**

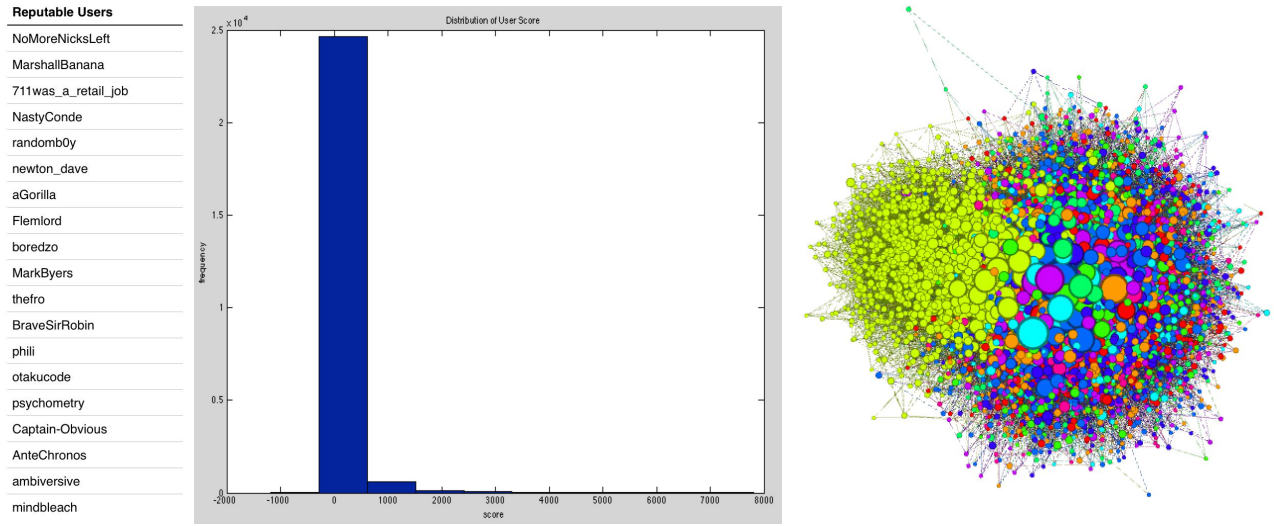
The data for the study is collected from the comments of the online community Reddit. Archive.org offers a archive of Reddit comments from October 2007 until May 2015 [4]. The 250GB dataset contains approximately 1.65 billion comments from more than 30 millions users; that is nearly every publicly available Reddit comment (~350,000 comments were not in the set due to API issues). Every single comment entry includes the author, timestamp, votes, topic and a number of scoring metrics implemented by Reddit. Due to the vast amount of data, we first run preliminary studies on the 2007 dataset and then focus on Subreddits over a number of years. The two studied Subreddits are r/relationships and r/books from 2007 to 2011; we think they are great candidates for the study because they each has a substantial number of subscribers and activities. Also they are more text-based, which help us conduct the text length analysis on the reputable users.

## RESULTS AND DISCUSSION

**Reputable User Labeling.** To identify the reputable users in Reddit, we compute a ranking score for every users and pick the top 20 from the sorted list of users. On Reddit, every comment has an upvote button and a downvote button. The number of upvotes and downvotes are captured in our dataset. Our ranking score for each user is the sum of net votes for the user: The distribution of the user score and the identified reputable users from the 2007 Reddit

$$Score(user) = \sum_{Comments(user)} upvotes - downvotes$$

dataset are shown in the figure below.



| Reputable Users |
|---|
| NoMoreNicksLeft |
| MarshallBanana |
| 711was_a_retail_job |
| NastyConde |
| randomb0y |
| newton_dave |
| aGorilla |
| Flemlord |
| boredzo |
| MarkByers |
| thefro |
| BraveSirRobin |
| phili |
| otakucode |
| psychometry |
| Captain-Obvious |
| AnteChronos |
| ambiversive |
| mindbleach |



**Figure 1: Reputable Users, the Distribution of User Score, and Graph Visualization Based on the 2007 Dataset**

The histogram in Figure 1 indicates that a vast majority of users have a low score. As the score increases, the amount of users decays almost exponentially. There are also a small amount of users with negative scores. Subreddits r/relationships and r/books from 2007 to 2011 also give a similar score distribution. The top users from Subreddits is shown below.

**Table 1: Reputable Users of r/relationships and r/books**

| r/relationships | | r/books | |
|---|---|---|---|
| AMerrickanGirl | ameoba | zem | Odusei |
| Sommiel | kouhoutek | GunnerMcGrath | cpt_bongwater |
| Skitrel | fddjr | munificent | omaca |
| drivebyjustin | fishwish | jordanlund | cojoco |
| priegog | sursurring | scottklarr | KR4T0S |
| Vinay92 | kornberg | epicpoll | camopdude |
| RagingErectus | mmmberry | fingers | judgebeholden |
| tothecore | ZorbaTHut | nycdk | punninglinguist |
| ylca | smacksaw | nista002 | AbouBenAdhem |
| slamare247 | sinxcosx | blackstar9000 | born_lever_puller |

**Social Graph and Visualization.** To visualize our 2007 data set, a social graph is constructed by representing each user as a node. An edge is added between two users when they both commented on the same thread. Our 2007 dataset results in a unweighted and undirected graph with 21,876 nodes and 5,446,372 edges. We then import the node and edge data in Gephi, identify and color different communities based on modularity, vary the node size based on the degree, and finally layout the graph using Yifan Hu's Multilevel option (see Figure 1). There are 17 identified communities. However, except for the community represented in green, all other communities seems to be inseparable from one another. The visualization does not convey a whole lot about the 2007 Reddit dataset.

For Subreddits, we take a different approach to generating the social graph. While each user remains represented by a node, a directed edge is added when a user responds to another user's comment; i.e. suppose user A responds to user B, an edge from A to B is added to the graph. Each edge is also given a weight, which is the number of responses from A to B. The r/relationships dataset results in a weighted and directed graph with 5,000 nodes and 16,210 edges, whereas r/books consists of 12,925 nodes and 59,519 edges. The graphs are visualized using Gephi with the Fruchterman Reingold layout option (see Figure 3).

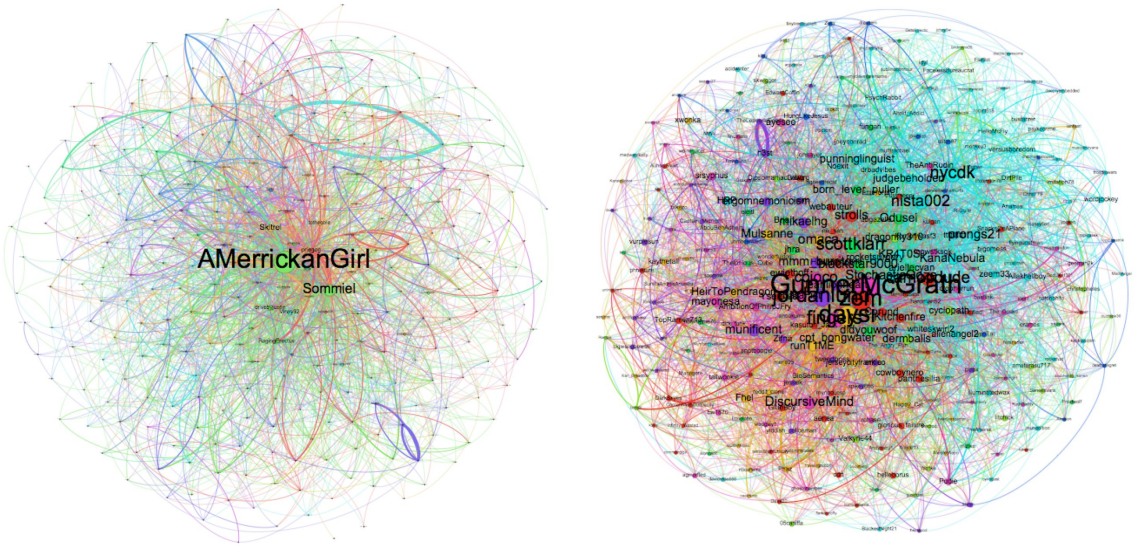



**Figure 2: Visualization of Subreddits r/relationships (left) and r/books (right)**

The Subreddit graphs provide a much clearer and less noisy visualization. They also reveal some interesting properties of the reputable users. For both r/relationships and r/books, we can find reputable users identified in Table 1 at the center of graph, which is also a gateway to different communities; we can infer that those reputable users have a high betweenness centrality. Also those users are in larger nodes, which indicate larger number of in-degrees.

**Network Property.** To see if we could identify reputable users using the social graph, we use each centrality metric degree, betweenness and closeness as a ranking score and extract the top 20 users based on the sorted results. The selected users for the 2007 data can be seen in

Table 2. We have limited success for the 2007 dataset. Each network property captures only 5 reputable users identified in Figure 1.

**Table 2: Top Users of the 2007 Dataset by Centrality Metrics (Bolded are reputable users determined by votes)**

| Degree | Betweenness | Closeness |
|---|---|---|
| **randomb0y** | **randomb0y** | **randomb0y** |
| deuteros | **boredzo** | deuteros |
| judgej2 | theDrWho | judgej2 |
| **boredzo** | malcontent | **boredzo** |
| db2 | cartooncorpse | db2 |
| bobcat | feces | bobcat |
| tony28 | db2 | tony28 |
| feces | deuteros | feces |
| sn0re | judgej2 | sn0re |
| cecilkorik | **NoMoreNicksLeft** | cecilkorik |
| **MarshallBanana** | **MarshallBanana** | Mr_Smartypants |
| **mindbleach** | soyabstemio | **MarshallBanana** |
| Mr_Smartypants | contrarian | **mindbleach** |
| **thefro** | Mr_Smartypants | redditcensoredme |
| redditcensoredme | **thefro** | **thefro** |
| rmuser | eshemuta | rmuser |
| malcontent | bobcat | malcontent |
| eshemuta | cecilkorik | eshemuta |
| reddit_user13 | redditcensoredme | reddit_user13 |

For the Subreddits, we have a directed and weighted graph. PageRank and HITS scores for each node and we use NDCG scores for evaluate the ranking. The weight used for each user in NDCG is the defined user score; i.e. the net votes for the user. The results can be seen in the table below:

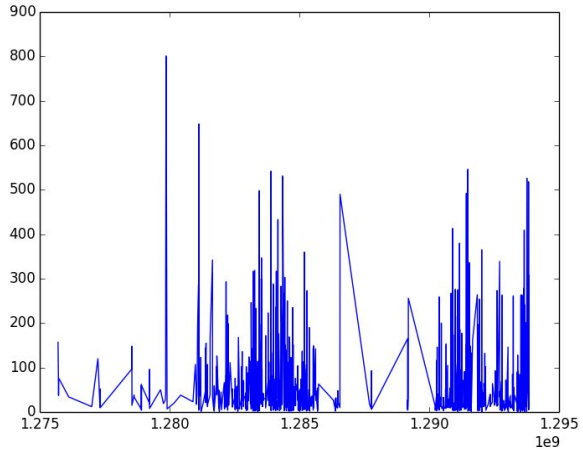
**Table 3: $NDCG_{20}$ Score and the Number of Top 20 Reputable Users Captured by Different Network Properties**

| | r/relationships | | | r/books | |
|---|---|---|---|---|---|
| Degree | 0.995 | 15 | Degree | 0.943 | 11 |
| In-Degree | 0.998 | 16 | In-Degree | 0.863 | 12 |
| Betweenness | 0.994 | 15 | Betweenness | 0.944 | 10 |
| Closeness | 0.993 | 14 | Closeness | 0.944 | 11 |
| PageRank | 0.997 | 17 | PageRank | 0.857 | 12 |
| Hub Score | 0.989 | 13 | Hub Score | 0.917 | 10 |
| Auth Score | 0.997 | 16 | Auth Score | 0.946 | 13 |

At a glance, network properties are much better indicators of the user reputation in subreddits. For r/relationships, the top users obtain high values of the computed network properties across the board. Properties such as high degree and betweenness are seen in the visualization in Figure 2. PageRank and In-Degree work especially well with a very high NDCG score and capture the most top users. Interestingly for r/books, the Auth score from HITS works best while PageRank and In-Degree are the worst performers in term of NDCG. It can be that the number of replies is less tightly coupled with the user's reputation for r/books.

**Text length analysis.** For this part we are interested in how top users write. We consider the length, or number of words, of each comment and try to see if top users write differently over time, and if top users write differently from other users.

First, we look at the difference in text length over time for 50 top users. For each top user, we fit a straight line with time of comment on the x-axis and text length on the y-axis, and test if the slope is zero. Below is an example plot for one of the top users. We try to fit a line using linear regression and then test if the slope is zero. The corresponding p-value is 0.09, indicating that we cannot reject the null hypothesis that the slope is zero, which aligns with what we see in the graph: there is no certain change in text length over time for this user.




For r/relationships, out of the 50 top users, 15 have a significant change in text length over time. 7 of them have a significant positive change, and the other 8 have a significant negative change over time. For r/books, the top users with positive and negative significant change is 5 and 4. From this we can conclude that the change in writing over time can be different for different top users, and most of them do not change.

We also tried comparing the average text length of top users and normal users. To do this, we calculated the average text length of top 5% users and also the average text length of the other users. We again test whether there is a significant difference between the two groups using Student's t-test. The p-value of this test is smaller than 0.05 for both Subreddits, indicating that the difference in text length is statistically significant for both Subreddits. Therefore, we can conclude that top users do write longer comments than the other users.

| Subreddit | Top users | Non-top users | p-value |
|---|---|---|---|
| r/relationships | 73.03 | 67.00 | 3.3e-2 |
| r/books | 40.17 | 34.86 | 4.4e-10 |

**How do users gain their reputation?** For this part we are interested in finding how reputable users gain their reputation. We assume there are two possible scenarios. The first scenario is reputable users gain reputation through a small fraction of their comments, which have much more upvotes than regular comments. The second scenario is reputable users gain reputation through a large number of comments which have more (but not much more) upvotes than regular comments. We can differentiate the two scenarios by looking at the variance of scores for each user. The variance of top users should be significantly larger than normal users in the first scenario, while the variance for different users should be similar in the second one.

Below is the average variance for different users in different Subreddits. We can see that there is a large difference in variance between different users for both Subreddits. To see whether this difference is significant, we can perform a t-test and see the corresponding p-value. The p-value is smaller than 0.05 for both Subreddits. Therefore, we say that the difference we see is significant. We conclude that top users gain their reputation through a few high-quality comments, rather than simply posting a lot of comments.

| Subreddit | Top users | Non-top users | p-value |
|---|---|---|---|
| r/relationships | 52.67 | 19.53 | 1.3e-3 |
| r/books | 178.07 | 32.27 | 2.6e-15 |

**Reputable User Detection.** We are interested in building a classifier for reputable user detection. We hope that this classifier can predict whether a user will be come reputable in the future based on the contents of its early-stage comments (maybe the first 5-10 comments).

One idea is to simply try to predict whether a user is going to be reputable based on his/her score for the first 5 comments. If the average score is above a certain cutoff, we predict that this user is going to be reputable. To evaluate whether this is a good predictor, we can look at the roc curve of this simple model. The roc curve of the two Subreddits are shown in Figure 3. We can see that this model actually performs well for r/relationships, but is only slightly better than random guessing for r/books. We think this difference is caused by the nature of the two reddits. For r/relationships, people with experience always provide more insights and therefore their comments are good from the beginning. However, for r/books, users may become reputable because of one book review, and this may not be detectable using their first 5 comments.

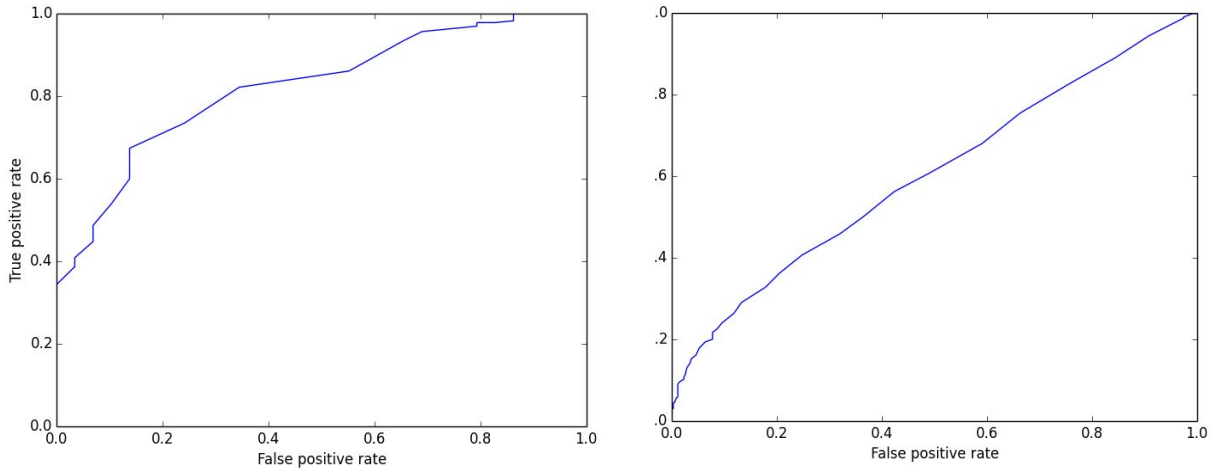



**Figure 3: The roc curve for r/relationships (left) and r/books (right)**

**Data modeling.** For this part we investigated what model can be a good underlying model for the data. We looked at several features of the graph to decide whether an existing model can be a good fit. We here discuss two potential underlying models, the Erdos-Renyi model and the preferential attachment model.

For Erdos-Renyi model, each pair of nodes independently connects with probability p. We consider estimating the value of p / q = p / (1 - p) from the graph in two ways and see if the estimations align with each other. First, we can estimate p using the fraction of edges in all possible pairs of nodes, and then estimate p / q using the estimation of p. We can get p / q = 0.023 using this approach. Second, we can look at different triads (tuples of three different nodes) that has two existing edges, and see if we have a third edge or not. We can divide the numbers of the two types of triads to get an estimate of p / q. Using this approach we get our estimate p / q = 0.515. We can see that there is a significant difference between the estimations of the two approaches, which means the Erdos-Renyi model is not a good model to describe the data. To be specific, we have much more small connected circles than a random graph, which is expected in an online social community.

For preferential attachment model, the degree distribution should follow a power law. Therefore we can look at the degree distribution of the graph to see whether we can use the preferential attachment model to describe the graph. The log-log plot of the degree distribution is shown in Figure 4 (left). We can see that the distribution is not as heavy-tailed as a power law distribution. Therefore the preferential attachment model is also not a good fit.

Since we did not find a model that can describe our data, we tried to do some modification to the models and see if we can build a model ourselves that fits the data we have. Our idea is that we can introduce a popularity for each node, which is fixed as the graph grows, and when a new

node comes, it chooses its connections using the popularity of the existing nodes. The log-log plot of the degree distribution of the simulated graph with average degree = 2 is shown in Figure 4 (right). We can see that the distribution is a better fit than preferential attachment.

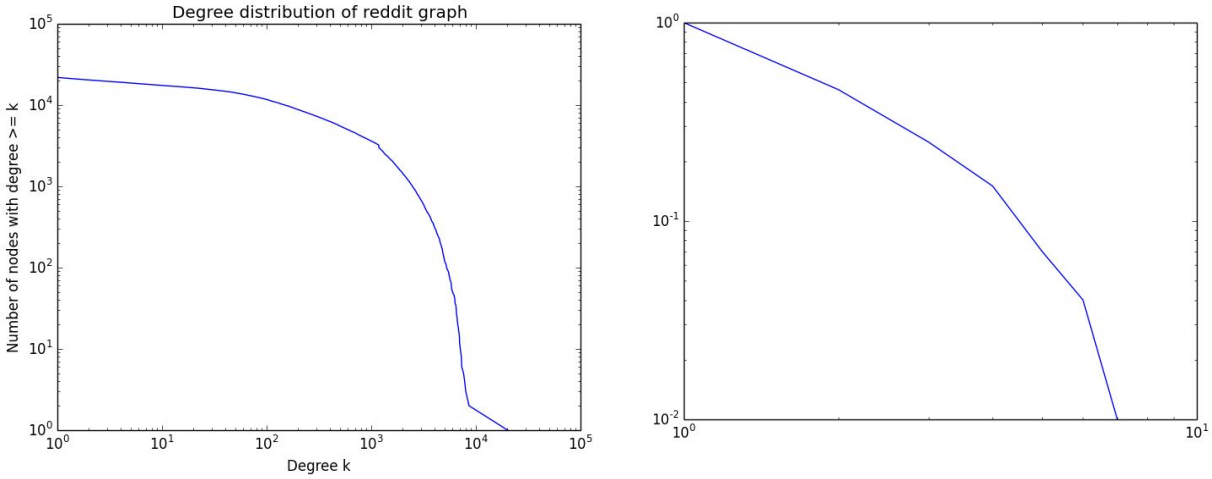



**Figure 4: Degree distribution for Reddit graph (left) and model simulation (right)**

## CHALLENGES

Since the dataset is extremely large, we focus on the 2007-2011 data in this report and plan to expand to other years in the final report. We utilized SQLite to avoid reading and parsing the data every time we run our study,. Through SQLite, we are able to store our dataset on a structured format and retrieve specified users and comments based on id.

## CONCLUSIONS

We proposed a metric to rank users' reputation and compared the ranking to the rankings of different centrality metrics. While the 2007 dataset does not exhibit a strong correlation between reputation and centrality, the Subreddits in our study show that the reputable users exhibit strong network properties, which are also identifiable from our graph visualizations. We also conduct a study on the writing behaviors of the top users. We learn that the change in writing over time is not consistent among the top users. However, based on our study, top users do write longer comments compared to other users. They also gain their reputation though a few high-quality comments, rather than earning it gradually through comments. We investigated the possibility to use a user's average score of the first 5 comments to predict whether this user will be reputable in the future. We found that this approach works well for r/relationships but does not work for r/books. We also propose a generative model including an intrinsic popularity for each node for the online community network for simulation studies.

**REFERENCES**


[1] Cheng, Justin, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "Antisocial Behavior in Online Discussion Communities." (n.d.): n. pag. ResearchGate. Web. 15 Oct. 2015.

[2] Mitzenmacher, Michael. "A Brief History of Generative Models for Power Law and Lognormal Distributions." Internet Mathematics 1.2 (2004): 226-51. Web.

[3] Zhang, Jun, Mark S. Ackerman, and Lada Adamic. "Expertise Networks in Online Communities." Proceedings of the 16th International Conference on World Wide Web - WWW '07 (2007): n. pag. Web.

[4] Baumgartner, Jason. “Complete Public Reddit Comments Corpus” Web. July 2015