Triaging Tweets: A network-structure based approach towards de-cluttering of twitter feeds

Akash Baid

1. INTRODUCTION

In several popular Online Social Networks (OSNs), the 'network feed' is a central component of the overall user experience on the social network platform. Important examples include Facebook, LinkedIn, Quora, and Twitter. In each of these online networking platforms, the network feed aims to provide a snapshot view of the recent developments about other users in each user's network. The type of information provided in the network feed varies from platform to platform, and can include network-level changes (e.g. announcements of new links, new nodes), and new content uploaded by the users.

Due to the growing number of friends/contacts that each user has on any OSN platform and due to the general increase in the amount of content (photos, videos, status updates, tweets, comments, etc.) created online, the network feed in almost all OSNs is becoming increasingly more cluttered with a deluge of information. The aim of this work is to understand this problem of network feed cluttering by analyzing a real-dataset for a specific OSN - Twitter. Our primary focus is on identifying distinct 'reasons' behind why a user might be receiving a large number of tweets in his/her feed. In other words, the target problem is that of triaging the tweets received by users by first classifying them into well-defined categories. The manner in which we can define categories is an open-ended problem and we focus on a network-structure based classification approach.

Two key techniques are proposed in this work to de-clutter the Twitter feed - (i) Thresholding for spam reduction, (ii) Grouping of tweets based on network-structure. The thresholding scheme is defined based on the insight received from analyzing a large-scale real-world Twitter dataset, which shows the presence of low-number of 'spammers' in many users' feed who are responsible for producing a disproportinally high number of tweets. We define a way to identify such users and then collapse their tweets, if the total number of tweets they post is more than a user-defined threshold level.

The second technique defined in this work, i.e. Tweet-grouping based on network structure is based on the premise that each user follows other users for several distinctive reasons, for example, because they are friends with another user, because they are an admirer of a celebrity, because a user has followed them, etc. While there are these different base-reasons behind following users, the tweets from all the people a user is following is currently shown in a single feed. We propose a mechanism to create different tabs or viewing panes to separate the tweets received from each class of users. In particular, we show the benefits of dividing the users that a given user follows based on two different classification types: (i) own-followers and non-followers, and (ii) friends, super-stars, and others. Lastly, we also show the possibility of defining a more structured way of forming a network from the tweets that a user receives and then applying known clustering algorithms based on this derived-network.

The rest of the paper is organized as follows: Sec. 2 provides details about the background and related works, while Sec 3 outlines the two main techniques that are proposed in this work. Next, Sec. 4 shows the insights gained from a large Twitter dataset, which is then used to analyze the proposed scehemes, as presented in Sec. 5. Sec. 6 concludes the paper and provides several different ideas for further work on this problem.

2. BACKGROUND AND RELATED WORKS

In [1], the authors argue that "finding interesting conversations to read is often a challenge, due to information overload and differing user preferences." The solution approach that they follow, however, is different from our approach. While we try to organize each user's feed to help them find useful content from within their feed, they propose a conversation recommendation system which takes into account the thread length, topic, and tie-strength, and presents converstaions from outside the user's feed. Similar insights about the problem are presented in several other user-studies; for example [2] shows that "most users would like to see more of what they care about, less of what they do not and more of who they are interested in, less of who they are not."

The most important example of triaging in the context of online social networks is Facebook's 'EdgeRank' algorithm [3]. While the details about the algorithm are not known in the

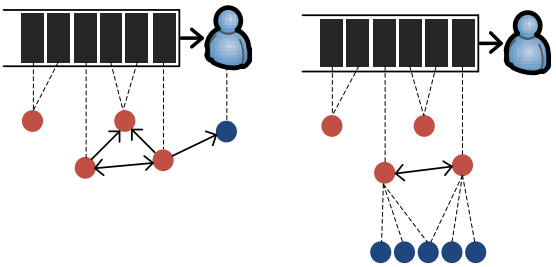

Figure 1: Forming a graph from each user's incoming feed. Two techniques are shown. *Left:* Authors of tweets that appear in the feed are connected by a link if they follow each other, *Right:* They are connected by a link if they have largely overlapping follower-base.

public-domain, the key inputs that this algorithm uses for selecting important updates from the set of all network updates received by user X are: (i) previous interactions between the author of a post and X, (ii) previous interaction between X and the port-type (photos, videos, etc.), (iii) reactions from users who already saw the post, (iv) amount of complaints or issues reported about the post. While this algorithm vastly reduces the amount of information shown to the users, the lack of transparency and lack of regard for the user's requirements are often criticized [4]. Some recent efforts have gone towards building a Facebook newsfeed selection algorithm that works outside the core platform[5], but such an approach is limited to small user-study groups.

The policy adopted by Twitter on the issue of feeds vastly differs from that of Facebook. While by default, Facebook only shows the most relevant items in the feed, Twitter displays every item in a simple reverse time-chronological order. We show that even when displaying all items, there could be ways to segregate the items in a meaningful manner. Several third-party applications, such as Jyst, TweetDeck, Twitterfall, and Tweets2d aim to change the way tweets appear to the users. However their focus is on the interface design and presenting easy to use tools to the user.

Due to the increasing importance of this problem, a few recent works have started to look at machine learning approaches to customize Twitter and other social network network feeds to suit individual needs [6, 7, 8]. While automatic learning of the users' likes and dislikes is a powerful technique for making the feed more useful for each user, we take a different approach in this paper: that of presenting the user with different types of classifications of the incoming feed items. Through this approach, the users retains the control of what they wish to see instead of being tied to the model that an algorithm builds for them.

3. PROPOSED APPROACH

The first part of our analysis focuses on finding more insight into the problem of bloated network feeds. In order to do this, we use a large Twitter dataset which was collected and analyzed for a previous study on trend characterization [9]. Sec. 4 provides the details about the dataset, and here we provide an overview of the key insights:

3.1 Underlying Problems

We identified two key factors that can help answer the following basic question: "why do some users receive a very large number of tweets in a given duration?".

- The presence of spammers: One of the main reasons why the number of received tweets per unit time can shoot-up for a user is the presence of specific spammers who post a large number of updates in a short duration.
- Following a large number of people: There is a high-degree of correlation between the total number of tweets received and the total number of user's followed. So users who follow a large number of other users tend to receive a large number of tweets per unit time.

The two factors do not work in isolation and we found users for which both problems existed to different extents. An interesting issue when analyzing the number of received tweets per user is to define the desired or allowed total number of tweets, i.e. a level above which the user's experience in terms of either information-gathering or entertainment-value starts to decrease. We take a parametric approach to this problem by allowing a tunable threshold for this. The issue of 'spam' classification is also a subjective one. In general if a user is posting a large number of tweets (say 10s per minute), a follower might be genuinely interested in getting their tweets or he/she might be annoyed by it. This requires a soft-approach for solving the problem, i.e. if the user so desires, it should be possible to view all messages, but the default view could be to curb the high-frequency posters in order to subdue their proportion in the feed.

For each of these two factors, we propose a simple solution. A simple thresholding scheme is proposed for tackling the high-frequency posters present in a user's feed. The basic idea is to set a user-defined threshold for maximum allowed number of tweets per-user per time-window (e.g. 60 tweets/month). The tweets that are above this threshold, are collapsed and displayed only when the user clicks on a small button below the last allowed tweet from the same user.

The following section explains the proposed technique for solving the 2nd problem in greater detail:

3.2 Network structure based grouping

The basic technique for de-cluttering the feed of users who follow a large number of people is to put some structure on their incoming feed. Several different approaches are possible here. For example, a useful classification technique could be to use a text identification mechanism which shows tweets on different topics in different viewing panes.

However we explore a network-structure based approach for classification. In particular we study four different types of classifications:

1. *Based on the overall in-degree of the author:* Group the

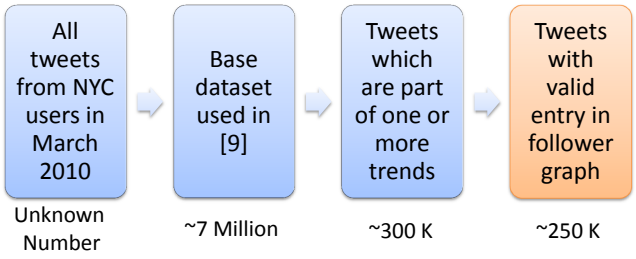

Figure 2: Overall data selection process.

authors of the posts present in a user's feed based on how many followers they have. The intuition behind in-degree based classification is to separate the tweets from celebrities and popular news channels from those posted by friends. Famous celebrities have a large follower-base on twitter while most friends of most users would have a low number of followers. We use two thresholds to define three categories: If the number of followers of an author present in the feed is below the lower threshold, their tweets are grouped into a separate tab (this might be labeled as 'Tweets from friends'). Similarly tweets from authors whose in-degrees are above the higher threshold are grouped into a 'Tweets from superstars' tab. The remaining tweets are shown in the main tab, for which the users have to do the triaging by themselves.

2. Based on the existence of reverse-direction links: Segregate the authors into two groups, one which are followers of the user under consideration, and the other which are not. People often follow back users who have followed them. But they are not necessarily interested in their tweets or at least not interested to the same extent as they might be in tweets from other users. Thus it might make sense to separate out the tweets from a user's followers from the main viewing pane and group them into a separate 'Tweets from followers' tab.

3. Based on the who-follows-whom graph of the authors: Find clusters of authors who follow each other and group their tweets. For people who follow a lot of their friends who also follow each other, it might make sense to use existing graph-clustering approaches to ascertain different clusters of friends that a user might have. Tweets from different clusters can then be shown under different tabs.

4. Based on common 'follower-ship' of the authors: A large overlap in the people who follow two different users indicates that some sense of commonality between them. In this case, a link is made between any two authors present in a selected user's incoming feed if more than a threshold percentage of their followers are common. Graph-clustering techniques can then be used to cluster based on this derived graph.

Fig. 1 shows how the main idea behind the formation of the last two graphs. Details about the performance of these approaches are presented in Sec. 5.

4. CLOSE LOOK AT TWITTER FEEDS

We used a large Twitter dataset consisting of more than 300,000 tweets and 61 Million users in this work. The dataset

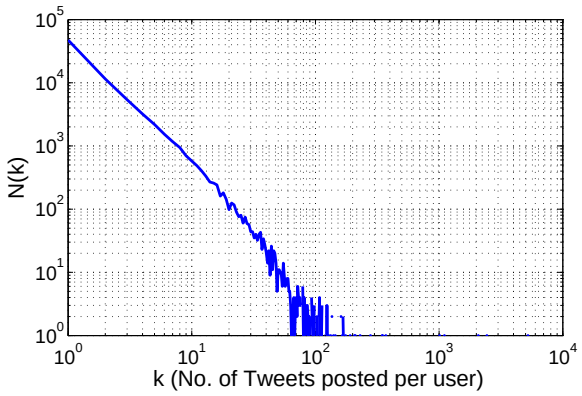

Figure 3: Distribution of the number of tweets posted per user (log-log scale)

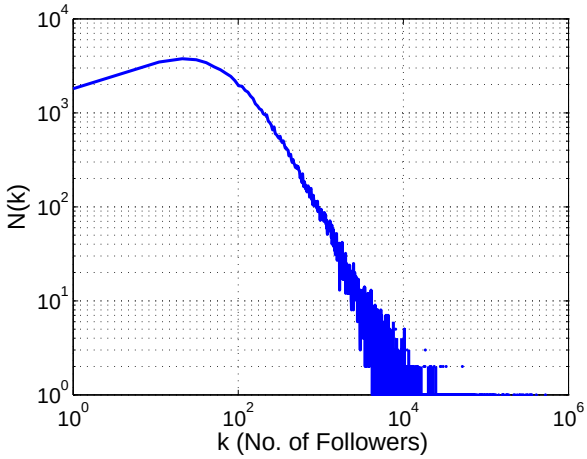

Figure 4: Distribution of the number of followers per user (log-log scale)

was used for a previous work on trend analysis [9], and as such only contains tweets that had one or more 'trending' terms as determined by the authors of that work. Fig. 2 shows the basic data-selection process showing the loss of a large number of tweets due to the process. In spite of the tweet dataset being incomplete, it can be used to present qualitative insights into the network feed characteristics.

The dataset from [9] consists of two parts: (i) Around 300,000 tweets posted by NYC based users in March 2010 which corresponded to 'trending' terms, (ii) The who-follows-whom graph of users that have posted anything that features in the set of tweets above. Figs. 3 and 4 show the basic characteristics of the two parts of the dataset.

By combining the who-tweeted-what and who-follows-whom graphs, we first derive a who-received-what dataset. Fig 5 shows the distribution of the number of tweets received per user for the top 1 Million users (in terms of the total number of received tweets). This shows that while most users only receive a few tweets in the selected month, there are some

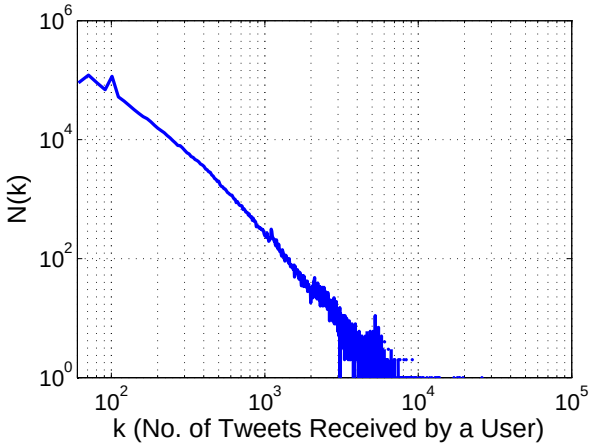

Figure 5: Distribution of the number of tweets received per user for the top 1 Mn users in a 30-day period (log-log scale)

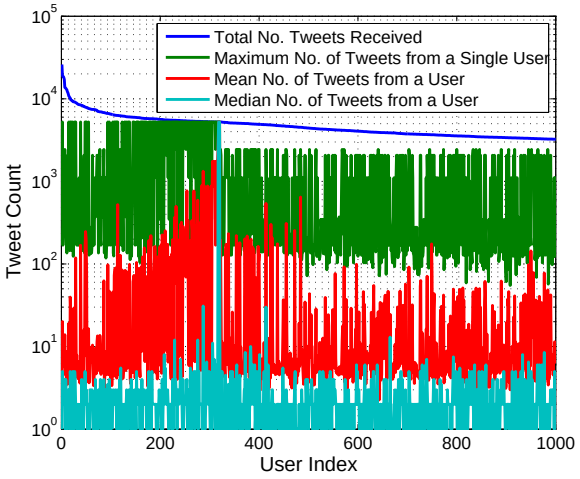

Figure 6: Total number of received tweets by the top 1000 users along with the max, median, and mean tweets per unique user.

users which receive a very high number of tweets. We found that even in our limited dataset, 18 users received more than 10K tweets in the 30 day duration. We selected the top 1000 users in terms of number of tweets received and use the tweets that they received for further analysis.

Fig. 6 shows the total number of tweets received by these top 1000 users along with the maximum, median, and the mean number of tweets received from a single user. The plots show that for many of the top users, a single author has posted a large fraction of the tweets in their feed. We found that a user with a Twitter handle: *mysalonbrand* posted 5,239 tweets during the 30 day period. However these high-frequency users are not the only reason why users feature in the top 1000 list as is clear by the difference between the

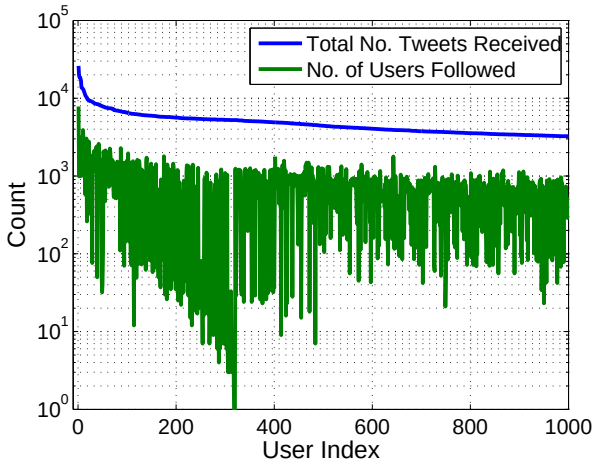

Figure 7: Relationship between the total no. of tweets received and the out-degree or number of following for the top 1000 users.

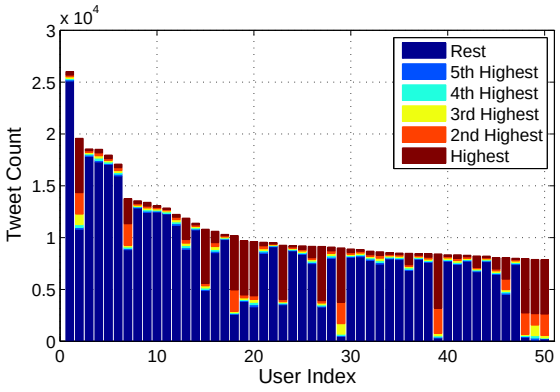

Figure 8: No. of tweets from the top 5 authors for each of the 1st 50 users.

total and the maximum lines in fig. 6.

The other underlying reason can be seen in Fig. 7 which shows the total number of tweets received along with the out-degree for each of the top 1000 users. The plot shows some correlation between the two for the the top few users. This is further highlighted in Fig. 8, which shows the breakup of the no. of tweets received from the top 5 authors, for the 1st 50 users from the previous figures. The presence of a large number of users for which the top 5 authors do not contribute to a large fraction of the total number of received tweets shows that the problem is beyond the presence of a select set of high-frequency users.

5. EVALUATION

5.1 Thresholding Result

In this section, we present the performance evaluation of the thresholding and grouping techniques described in Sec. 3. Since both the techniques remove a certain number of tweets from the main feed of the user (either collapses them or

Threshold (per month)	No. of Users with tweets suppressed	% of total tweets suppressed
60	150	7.05
150	24	4.84
300	8	4.14
1500	3	2.07

Table 1: Different threshold values and corresponding results

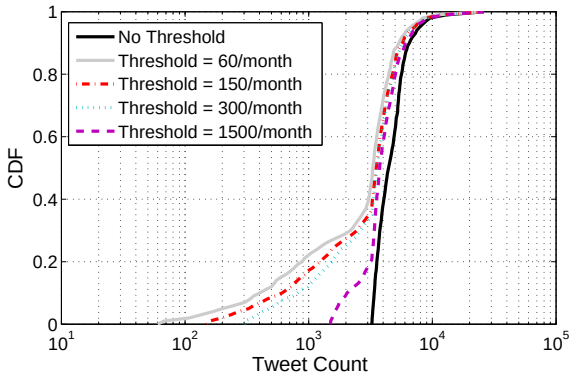

Figure 9: Distribution of the number of tweets in the original tweet and the corresponding values when using four different thresholding values.

moves them to other tabs), the important thing to measure is the number of tweets that are still left in each user's feed.

We use 4 different thresholds in terms of tweets per month. Table 1 shows the values used and the resulting percentage of users affected. Fig. 9 shows the effect of thresholding in the number of tweets remaining in the feed. Clearly, for the bottom 40% of the users in the CDF, thresholding forms a easy-to-use knob to adjust the amount of tweets they see. But the top 60% of the users are relatively unaffected by this technique.

5.2 Grouping Results

Fig. 10 shows the distribution of the number of unique authors present in each of the top 1000 users' feed. This would determine the size of the graph, and from the plot we can see that the values range from just a few node to a few-thousand nodes.

5.2.1 *In-degree based Classification*

Fig. 11 shows the two different thresholds we used for the in-degree based classification. The 2nd set of thresholds is a more aggressive setting in which a larger percentage of the users are classified into either 'friend' or 'superstar' categories.

The performance of the grouping scheme is shown in Fig. 12, which shows the total number of tweets remaining in the un-classified tab for each of the top twenty user for the two threshold values mentioned above. This shows that a large fraction of the tweets in a user's feed can be separated out to other tabs based on this method of classification. Moreover,

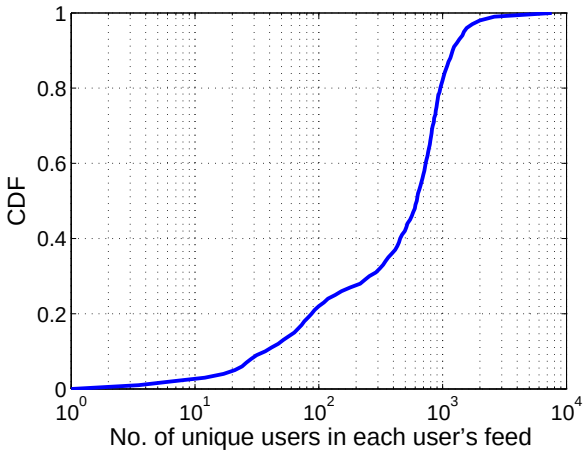

Figure 10: CDF of number of unique authors in each users feed for the top 1000 users.

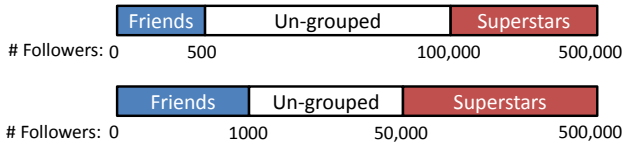

Figure 11: Thresholds used for classifying authors based on their in-degrees

the user can select the level of aggressiveness by selecting the thresholds.

5.2.2 *Follower/Non-follower Classification*

Fig. 13 shows the fraction of tweets received from followers and non-followers by the top 75 users (with out-degree information available in the dataset). In our analysis we found that 61% of all authors who appear in the feeds of these top users are followers of the corresponding users. Also, 42% of all tweets are from followers. Thus if a separate 'Tweets from followers' tab is made, then 42% of the tweets can be removed from the main viewing pane.

5.2.3 *Other graph-based approaches*

The other two graph-based approaches, which are: (i) grouping the authors present in a given user's feed based on their who-follows-whom graph, and (ii) grouping the authors present in a given user's feed based on common followers graph show a lot of potential but could not be analyzed in full in this work. Part of the problem in these two approaches is the sparseness of the resulting graphs. Figs. 14 and 15 show the number of nodes and edges in the graphs formed from each of the 150 users with least value of the number of unique authors. For users with very high number of unique users, creating these two graphs requires a non-trivial amount of computations.

6. CONCLUSIONS AND FUTURE WORK

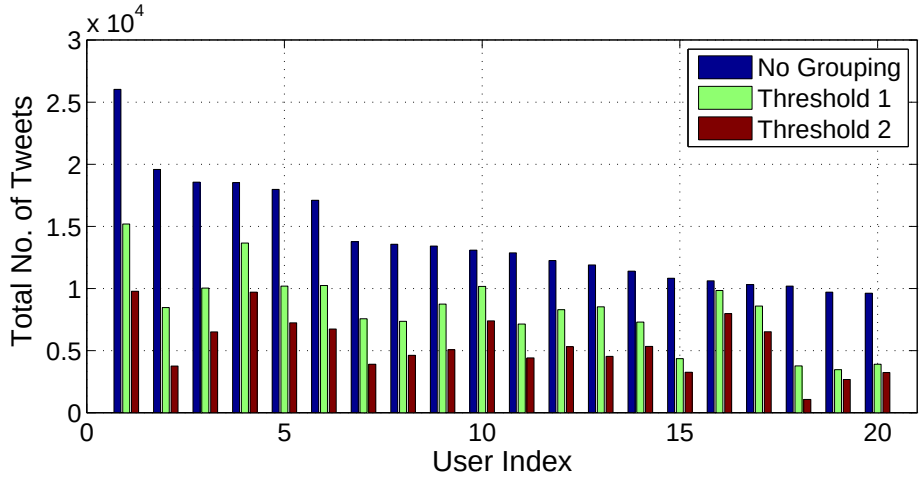

Figure 12: Performance of the in-degree based classification scheme: No. of tweets from un-classified authors shown for the two threshold-set used.

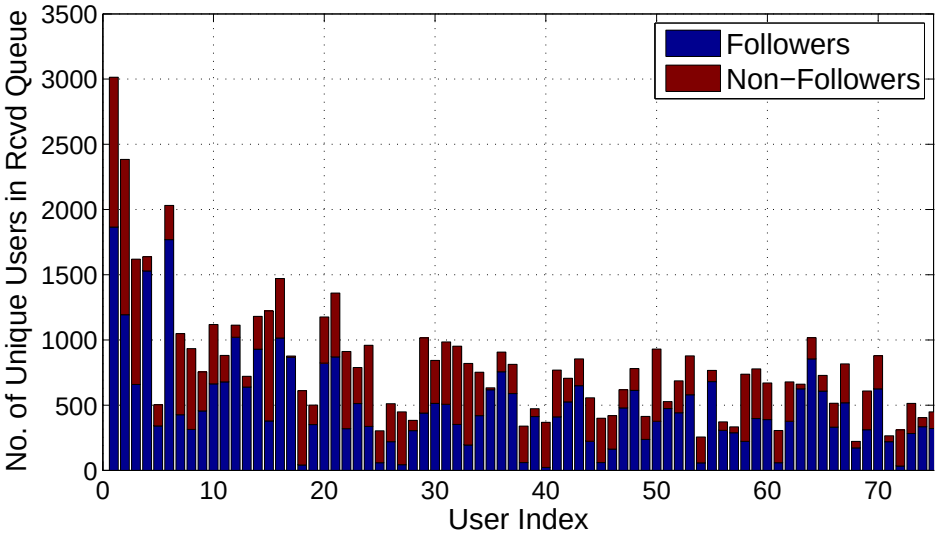

Figure 13: Performance of the follower/non-follower classification scheme: No. of tweets from each group is shown as fractions of original.

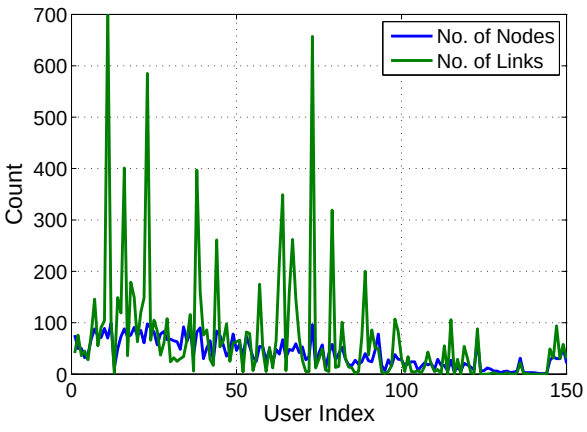

Figure 14: Number of nodes and edges from the who-follows-whom graph of the authors present in each user's feed.

Social network feeds for most users increasingly contain a very high number of items. This is a challenging and involved problem which has not received much importance from the academic community. While there are many subjective and general policy related issues that have to be thought through, in this work, we suggested a set of simple algorithmic techniques to mitigate the basic problem. The two solution approaches: thresholding and grouping

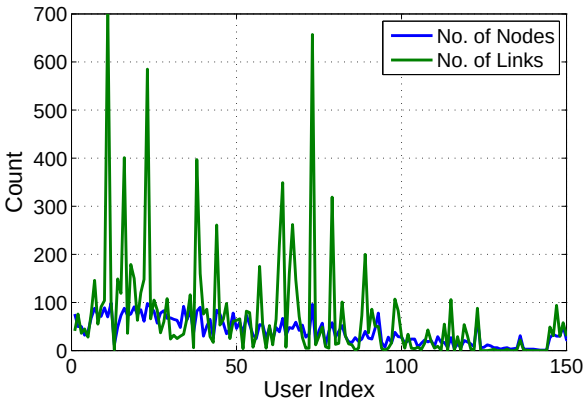

Figure 15: Number of nodes and edges of the graph formed by linking authors with overlapping follower-base.

targets high-frequency authors and high-degree users respectively. We explore several different network-structure based approach for segregating the incoming tweets in a user's feed, and compared the performance through a large Twitter dataset.

Further work on this problem requires a more complete dataset which must contain all the tweets received by the users present in the dataset. The two clustering based approaches outlined in Sec. 3 requires additional tricks to work with the sparse graphs. Combination of several different approaches would lead to better results and needs to be tested on a larger dataset. It is also important to consider the ease-of-use of such a system since a classification criteria which is too obtuse for the user might result in making the task of the user more complex instead of simplifying it.

7. REFERENCES

[1] J. Chen, R. Nairn, and E. Chi, "Speak little and well: recommending conversations in online social streams," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, 2011, pp. 217–226.

[2] K. Y. Lee and J. L. Hong, "A user survey on search ranking algorithm for social networking sites," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, 2012, pp. 995–999.

[3] J. Kincaid, "EdgeRank: The Secret Sauce That Makes Facebook's News Feed Tick," http://techcrunch.com/2010/04/22/facebook-edgerank/.

[4] N. Bilton, "Disruptions: As User Interaction on Facebook Drops, Sharing Comes at a Cost," http://bits.blogs.nytimes.com/2013/03/03/disruptions-when-sharing-on-facebook-comes-at-a-cost/, March 2013.

[5] T. Paek, M. Gamon, S. Counts, D. M. Chickering, and A. Dhesi, "Predicting the importance of newsfeed posts and social network friends," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, 2010, pp. 1419–1424.

[6] M. Honkala and Y. Cui, "Automatic on-device filtering of social networking feeds," in *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, ser. NordiCHI '12, 2012, pp. 721–730.

[7] S. Berkovsky, J. Freyne, and G. Smith, "Personalized network updates: increasing social interactions and contributions in social networks," in *Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization*, ser. UMAP'12, 2012, pp. 1–13.

[8] S. Berkovsky, "Network activity feed: finding needles in a haystack," in *Proceedings of the 4th International Workshop on Modeling Social Media*, ser. MSM '13, 2013, pp. 1:1–1:1.

[9] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on twitter," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 5, pp. 902–918, May 2011.