Detecting Organizational Accounts from Twitter

Based on Network and Behavioral Factors

by

Chinmay Chandrashekhar Gore

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved June 2017 by the
Graduate Supervisory Committee:

Hasan Davulcu, Chair
Ihan Hsiao
Baoxin Li

ARIZONA STATE UNIVERSITY

August 2017

ABSTRACT

With the rise of Online Social Networks (OSN) in the last decade, social network analysis has become a crucial research topic. The OSN graphs have unique properties that distinguish them from other types of graphs. In this thesis, five month Tweet corpus collected from Bangladesh - between June 2016 and October 2016 is analyzed, in order to detect accounts that belong to groups. These groups consist of official and non-official twitter handles of political organizations and NGOs in Bangladesh. A set of network, temporal, spatial and behavioral features are proposed to discriminate between accounts belonging to individual twitter users, news, groups and organization leaders. Finally, the experimental results are presented and a subset of relevant features is identified that lead to a generalizable model. Detection of tiny number of groups from large network is achieved with 0.8 precision, 0.75 recall and 0.77 F1 score. The domain independent network and behavioral features and models developed here are suitable for solving twitter account classification problem in any context.

To my parents, family and friends...

ACKNOWLEDGMENTS

I would like to thank Professor Hasan Davulcu for guiding me throughout the process. Thanks to him, I was able to work at CIPS lab and feel privileged to be part of a great team and amazing projects. I would also like to thank Dr. Baoxin Li and Dr. Ihan Hsiao for supporting me through this research and being part of my thesis committee.

I would like to thank my colleagues Pankaj, Kallol, Amin, Mert, Sultan and Nyunsu for helping me out. This would not have been possible without your support.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Last decade has seen a revolution in online social networks. With the evolution in electronic industry and the invention of the Internet, social media access became accessible to large mass of population. Online social networks has become integral part of every persons lifestyle where people leave footprint of their thoughts and actions. Twitter is one of such an online social network which allows users to publish text messages with limited length (140 characters). As per Twitter official site, there are 31 million active users on monthly basis and 80% of those are outside of United States. Twitter supports more than 40 languages. Twitter allows users to track their friends activities by following them. Users can also retweet and comment to the tweet messages by others. Since the Twitter data is publicly accessible through APIs, it is the favorite domain for social network research.

While other social networks like Facebook, Google+ etc allow creating user groups, Twitter does not have such feature. People usually use identical hashtags to talk about various common issues on twitter. Institutions and organizations such as universities, companies and government offices make use of the Twitter platform to announce policies and real-time updates. Organizations also make use of these groups to advertise their products and communicate with mass populations. These twitter accounts usually have a common theme and people communicate through these Twitter pages by retweeting and retweeting each other. In this paper we study such groups which are linked to political organizations and NGOs.

There has been a lot of work done on community detection in twitter. The traditional community detection algorithms make use of spectral clustering to partition

the graph into various communities. The main problem there is to find latent groups that are formed by analyzing the interaction amongst people. The ground truth for such communities is found by gathering the actual groups that are formed on social media and assigning all the users from same group into one community. However Twitter does not have the notion of groups like Facebook or Youtube. On Facebook, people can create groups with their families, close friends or events. On Youtube, people can subscribe to various channels which serve the purpose of groups. The posts made in a group are broad-casted to entire sub-community and these bind the like-minded people together. In this study, the main focus is on twitter handles which are official pages of political organizations and NGOs on Twitter. The fact that there are very small number of these accounts makes it a difficult task. The task is similar to twitter bot detection where the percentage of bots is very small. However, the behavior of the groups is not drastically abnormal from the entire population as compared to automated bots and thus it is important to gather a set of features that can easily discriminate between group vs non-groups. To our knowledge this is first kind of work that aims to detect the official organizational accounts on Twitter.

The rest of the thesis is organized follows -

Chapter 2 gives summary of similar work that has been done. Chapter 3 describes the Bangladesh dataset that is used and ground truth collection process. Chapter 4 is analysis of the features of Bangladesh tweets that can discriminate between organizational vs non-organizational accounts. Chapter 5 presents the analysis and results presented in the data. Finally Chapter 6 concludes the thesis with pointers to future work.

Chapter 2

RELATED WORK

To best of our knowledge, this is the first kind of work that focuses on finding organizations accounts on twitter. In this section, summary of similar work is presented. The closest to the problem explored in this research is work done by Wu *et al.* (2011). It focuses on dividing the twitter accounts into different categories and then analyzing the flow of information between those. Their main focus is on validating the two step communication flow model. The paper divides users into elite and ordinary users. The users are divided into 4 types - media, celebrities, organizations and bloggers. In the first step, they handpick the representative accounts for each of the four categories and crawled the content of those accounts. Then a set of discriminating keywords was manually generated for each category. A score is then calculated for an unseen account based on its tweets and is then classified into corresponding category. Next the elite set of users are extracted from each of the categories by utilizing their frequency in the twitter lists.

Since the proportion of organizational accounts is very low, study of anomaly detection on Twitter is studied. Bot detection from twitter network is very similar to finding organizational handles on twitter since the fraction of organizational handles from a large network of individuals has to be performed. There has been a lot of work done in the area of bot detection in Twitter Chu *et al.* (2010) and Subrahmanian *et al.* (2016).

Chu *et al.* (2010) propose an approach for detecting bots by generating an entropy measure based on time series features, probabilistic features based on the textual corpus and account properties of user such as description, URLs. Finally, the decision

maker combines these features into one single model to produce the result.

The DARPA 2015 challenge Subrahmanian *et al.* (2016) focuses on detecting the specific kind of bots from the pro-vaccination topic on Twitter. These bots try to influence the behaviors of the community through spreading a particular sentiment and thus can be used in negative manner. Six teams that participated created a list of syntactical, network and temporal features that they used to classify the influence bots over others from a set of 7K user accounts. There has been recent work by Yu et al Survey on social media anomaly detection that summarizes the work that has been done to detect the point and group anomaly patterns on social media. The work used graph and activity based information of the users to detect anomalies.

Rao *et al.* (2010) propose a method to classify latent user information of the users e.g. age, sex, political affiliation and regional origin. The paper focuses on textual features of the user attributes to find patterns in tweeting that can detect users latent attributes. This method although effective, it is not suitable for multi-language domain that we use for our dataset.

Recent work by Varol *et al.* (2017) works on bot detection with around thousand features. The features are divided into following categories - user-based, friends, network, temporal, content and sentiments features. However, in this thesis the aim is to build the language independent organization detector, so most of the features cannot be used here. According to them at least 9% of the total accounts are bots. In our case the percentage is even lower than that. Varol *et al.* (2017) proposes new model that improves the recall for bot detection problem.

The earliest work on social networks by Mislove *et al.* (2007), finds interesting characteristics of online social networks. It shows the scale-free nature of the OSNs by analyzing the Flickr, LiveJournal, Orkut and YouTube networks. All the networks follow linear trend on log-log scale and have strongly connected core, in contrast to

the web graph.

Another excellent paper by Zuber (2014) presents a detailed survey of data mining techniques for social networks.The paper covers most of the historical and recent techniques for classification, semi-supervised approach and clustering on social networks.

Chapter 3

DATASET

The dataset consists of all the tweets from Bangladesh during 5 month period from 1st June, 2016 to 31st October, 2016. Twitter GNIP API was used to collect all the tweets from location originating from Bangladesh. Some of the tweets that are not geotagged i.e. when the users dont check in the location. These tweets are marked for their geolocation using GNIP location prediction. This algorithm involves predicting the user location based on their IP addresses and other such attributes.

Number of tweets	7090560
Number of users	150000
Minimum timestamp	June 1, 2016
Maximum timestamp	October 31, 2016
Tweets Location	Bangladesh
Languages used	English, Bangla

Table 3.1: Tweets Dataset

3.1 Ground Truth Collection

Since the user accounts for the collected tweets are not labeled, we manually created ground truth for labeling the user accounts. The user accounts were divided into following categories:

(1) Celebrities

(2) News

(3) Political Organizations

(4) Entertainment groups

(5) NGOs

(6) Individuals

The remaining users were marked as unknown. The ground truth was collected using handpicking and searching for keywords for each of the account types in the corpus. Some of the keywords were borrowed from Wu et al. All the nodes were sorted in descending order of their pagerank centrality and degree centrality. Top 200 of these two lists was manually labeled. In addition to these following method was applied to label each of the categories.

Labeling news: Created a list of popular TV channels and newspapers in Bangladesh. Using this list, the twitter handles for each of the channel/newspaper was found from corpus.

Labeling celebrities: An approach similar to new labeling was used to label celebrities. Collected a list of movie actors, politicians and public figures from Bangladesh and searched for their twitter handles in the existing corpus. Some of the celebrities was found using tracking the number of followers. Found top followed accounts and manually verified account type of each.

Labeling political organizations: Collected a list of political parties and their respective handles from ASU political domain experts from school of religion. This list was further expanded upon by going through top friends of these accounts. In addition to that keywords specific to parties were used to search for these accounts in the corpus. We also scanned through wikipedia pages of each political organizations in bangladesh and found corresponding handles for those.

Labeling NGOs: NGO labelling was done mainly based on keyword search. Some of the ngos were found by looking for local branches of globally active organizations e.g. red cross.

Labeling Individuals: After going through the user descriptions, found patterns of in user profile description. Using these patterns, created a list of regular expressions that find certain phrases like Im, I like to etc. Using these regex we could label approx. 30K individuals.

The table below shows the distribution of dataset after labelling the data.

Account type	Count
Individual	30957
Celebrities	17
NGOs	68
News	62
Political organizations	35

Table 3.2: Ground Truth Data

3.2 Collecting Followers

The gnip data does not provide the follower information of the collected twitters. The followers graph however is very crucial since higher followship means higher publicity. Hence the followers of all the users was essential to be gathered.

Twitter public API is very limited in sense that it allows only 15 requests every 15 min. So it is not possible to gather the followers information of all the 150K users that we have. So, we concentrated on gathering the followers information of all the labeled users i.e. users in the ground truth. We created 25 crawlers to gather the

data over period of 4 weeks. We gathered only 1 million users of each users and threw away rest for practical reasons. The collected data contained 12 million users.

Chapter 4

FEATURES

In this section we describe the features that were successfully used to discriminate between organizational accounts and other. These features are language independent i.e. they rely only on non-textual features of the users.

There has been a lot of work done in the area of social network. The social networks differ from normal networks by following properties

(1) Scale free: Barabási and Albert (1999) showed that a large number of empirical networks are scale free i.e. they follow power law distribution. This behaviors is seen in most of the social communities, shown later by Mislove *et al.* (2007). Same analogy can be seen in economy. Most of the money goes to top 10% of the society and rest 90% share remaining resources. This behavior was observed in a large set of networks - WWW, Social networks and biological networks. This is true for twitter networks as well.

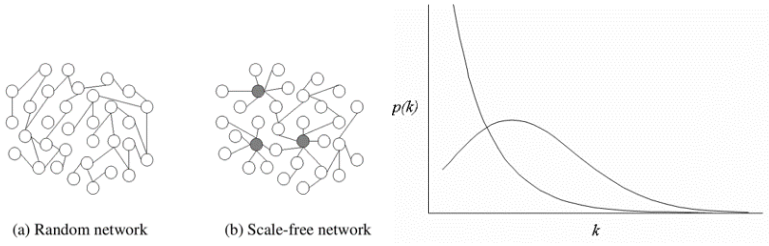

Figure 4.1: Degree Distribution of Random Network and Social Network

If we were to construct a random network, the degree distribution of that network would follow normal distribution. But, for social networks, the distribution is power law. According to Barabasi, it is because of the preferential attachment of the nodes

- incoming new nodes are more likely to pair with high degree nodes in the network.
(2) Small world: The world is indeed small! This was proved by Travers and Milgram (1969). Even if there are millions of ways to reach from Nebraska and Boston to Massachusetts, the number of hops made by each of the 64 letters that reached final destination is not a big number. In fact average diameter of the world graph is estimated to be 7. Here diameter means length of longest shortest path. For Facebook and Twitter the average path length is 4.

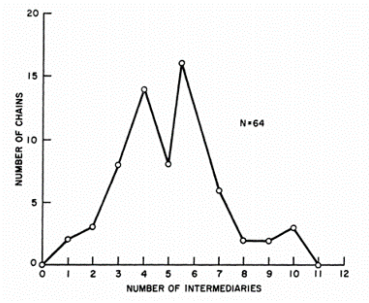

Figure 4.2: Small World - 6 Degrees of Separation

(3) High clustering coefficient Compared to random graphs, the clustering coefficients of the social networks is quite high. Clustering coefficient is ratio of the triangles that are present in the graph. It is ratio of how many friends are connected to each other vs total possible friends interconnections.
There are two types of clustering coefficients global and local. Global clustering coefficient is ratio of number of triangles in a graph by ratio of possible number of triangles. This is called transitivity. Local clustering coefficient is for each node and it is calculated by fraction of friends of friends in for a node.
Following 3 networks were created from the tweet corpus:
(1) Retweet network: From the tweets that are collected, 28% of the tweets are retweets. For each of the retweet by a user, we add link from the user to original author of the tweet. This creates a directed weighted graph where original author

receives an incoming edge and the weight of the edge represents how many times this author was retweeted. The table shows the characteristics of retweet network.

Nodes	308,477
Edges	681,404
Number of connected components	4,305
Average node degree	75
GCC Nodes	299,890
GCC Edges	675,682

Table 4.1: Retweet Network Stats

(2) User mentions network: This network is generated by adding edge between the users that mention each other. Again the direction A to B indicates the user B that was mentioned and A represents the mentioned. The weight represents how many times the user has been mentioned.

Nodes	335,678
Edges	431,437
Number of connected components	11,755
Average node degree	48
GCC Nodes	298,337
GCC Edges	405,813

Table 4.2: User Mentions Network Stats

(3) Followers Network: High followship is a strong indicator of a highly popular or influential user. The GNIP data however does not provide the followers information of the users. We used twitter API to crawl the followers of each of the users in the

dataset. A followers graph was built using these.

Nodes	12,604,797 (12 million)
Edges	25,942,312 (25 million)
Number of connected components	1079
Average node degree	1078
GCC Nodes	12,600,824
GCC Edges	25,939,414

Table 4.3: Followers Network Stats

In the following sections, network and behavioral features of the users are presented. Each feature is followed by a scatter plot of the users and groups.

4.1 Network Features

We follow following notation for the social network graph.

User Node	V
Graph	G
Edge	E
Number of nodes in graph	N_v
Number of edges in graph	N_e

Table 4.4: Notation

4.1.1 *Degree Centrality*

Degree centrality of node V is fraction of nodes it is connected to. This is further divided into in-degree and out-degree. We use normalized degree centrality.

$$C_V = \frac{deg(V)}{\sum_{i \in G} deg(i)}$$

Similarly we can define in-degree and out-degree of every node. Following plot shows distribution of degree centralities for retweet, user mentions and followers graph. Here the dots (O) are marked as normal users and the groups are marked by cross (X).

For celebrities, we observe high in-degree centrality and low out-degree centrality.

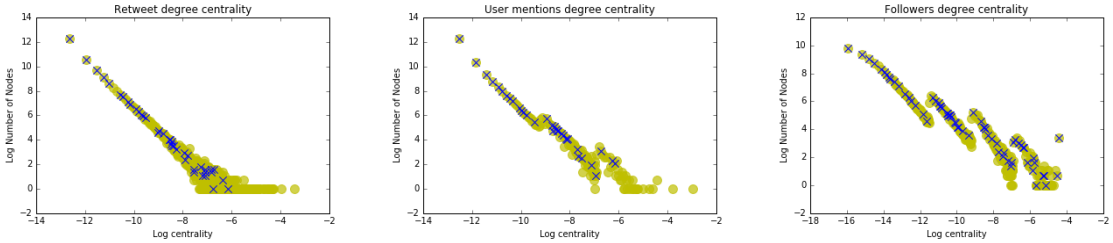

Figure 4.3: Degree Centralities

Which asserts the fact that while celebrities are followed by large population, they follow very few users. Similar trend is seen for news accounts.

Groups are generally seen to be on the head of the power law distribution. For groups user mentions out centrality is low which might be because the groups are mentioned more than they mention others.

4.1.2 *Pagerank Centrality*

Pagerank centrality (Page *et al.* (1999))is an extension of Eigen value centrality. It is the most popular type of centrality since invention of Google by Larry and Page. It was used to rank web documents from a directed graph of web pages.

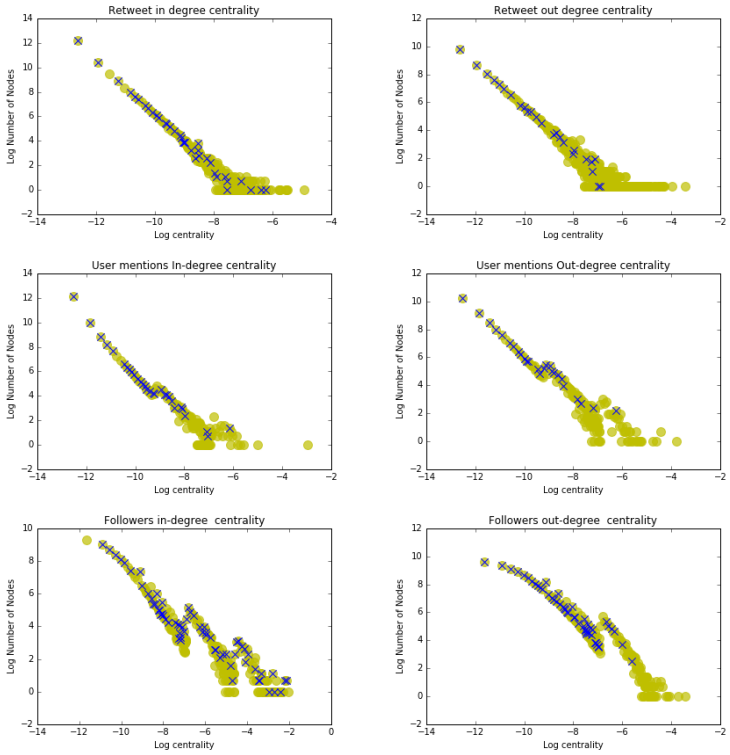

Figure 4.4: In-out Degree Centralities

Pagerank is usually computed iteratively. We compute the pagerank using damping factor of 0.85 and the max number of iterations 100. Following were the values observed for pagerank of retweet, user mention and followers graph. We can see that

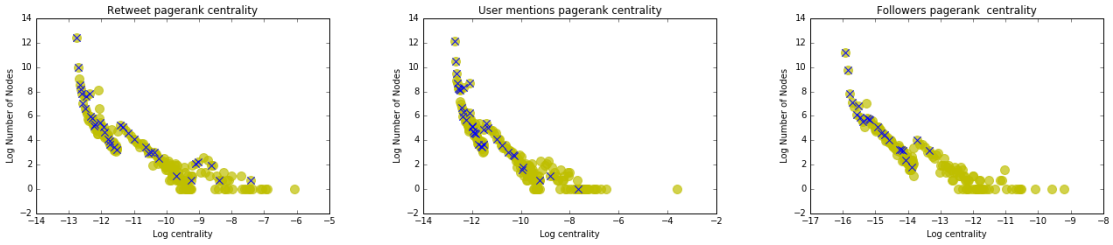

Figure 4.5: Pagerank Centralities

groups usually have low pagerank in followers and user mentions network. Retweet has more outliers.

4.1.3 K-core Centrality

K core centrality (Seidman (1983)) is another important centrality measure. It is computed by recursively pruning the nodes till none remain. In ith iteration, we progressively remove nodes with degree i. So for the first iteration, we remove all the nodes with degree 1 till there are none with degree 1. All these nodes are assigned kcore degree of 1. The graph will then have minimum degree of 2. We repeat this procedure until there are no nodes in the graph. We can see for the groups, values

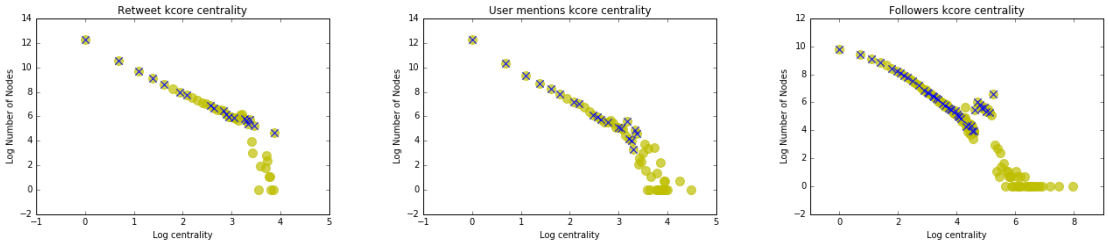

Figure 4.6: K-core Centralities

are clustered together in first half of the graph, separated from other accounts.

4.1.4 Clustering Coefficient

Clustering coefficient(Watts and Strogatz (1998)) for node V is defined as the ration of actual number of triangles around that vertex to possible number of triangles.

$$CC_V = \frac{2T(V)}{deg(V) \times (deg(V) - 1)}$$

where $T(V)$ is the number of triangles around node V

Clustering coefficient represents number of friends that are friends of each other. If all the friends are connected to each other then the clustering coefficient is 1 if no friends are connected to each other coefficient is 0. Our hypothesis is that the friends connectivity should be different based on the type of account. For example, celebrities

have diverse set of followers which are not connected to each other. Following graphs show clustering coefficients for the Bangladesh users.

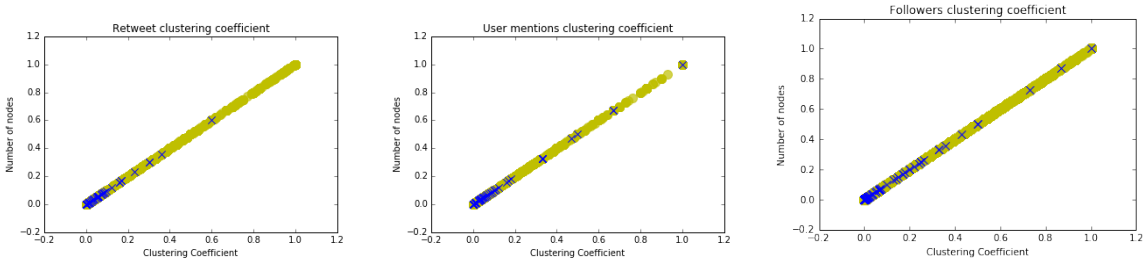

Figure 4.7: Clustering Coefficients

From the plot we clearly see that the groups have low clustering coefficient in general in all three graphs. We can see some outliers who have very high clustering coefficients.

4.2 Temporal Features

In addition to network features of the users, we analyze the temporal and spatial features of the users that deal with tweeting patterns and locations of the users.
For every user, we have a bunch of timestamps at which the user tweeted. We did not find any patterns in the timings of the tweets. Most of the users are active throughout the day and less number of them at night. Also there is no clear distinction between the timings of tweets and type of user account.
We however found patterns in tweeting patterns of the users. We calculated the variance of tweet timestamps of each user and plotted it. We calculated the variance by calculating number of seconds from minimum timestamp we had and dividing it by 5 months seconds to normalize it. We then calculated the variance and information gain from these.
While entropy of groups is clustered around the center, variance of groups is clustered around the beginning.

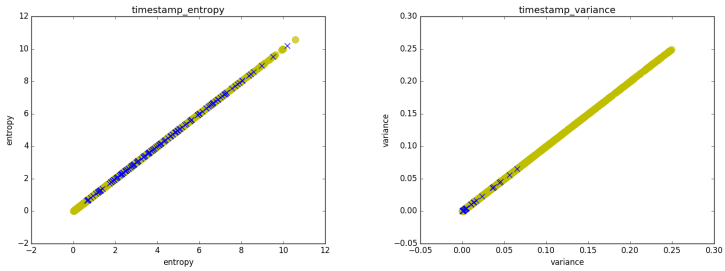

Figure 4.8: Temporal features

4.3 Spatial Features

From the data that we collected we have a small number of tweets that contain the geo location of the tweets. Individuals when they move, are going to check in from different locations and we can catch that. The hypothesis is that we can catch the individuals and celebrities that use cell phones to check-in to twitter.

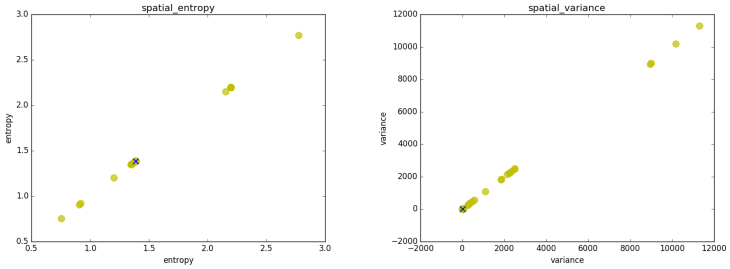

Figure 4.9: Spatial Features

We have extremely small number of accounts that have check-ins and even fewer for the accounts that are groups.

4.4 User Profile Features

For all the users in database we gathered their profile information such as user description, favorites count etc. There has been a lot of work done in this area which

uses profile based features to classify the users based on their profile pictures, color of the background/theme etc. We use following features for our classification task.

4.4.1 Followers to Friends Ratio

For celebrities, we expect a large number of followers and less number of followees (friends). We plot this ratio for the all types of accounts.

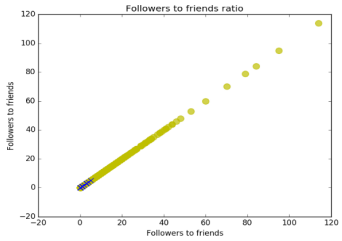

Figure 4.10: Followers to Friends Ratio

We found that the groups have less friends to followers ratio.

4.4.2 Favorites Count

Twitter allows users to like any tweets and this information can be captured in favorites count. Assumption is that the individual users should have more favorite count on average than groups.

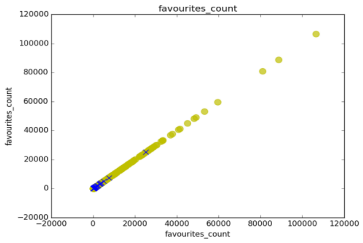

Figure 4.11: Favorites Count

4.4.3 *Listed Count*

Twitter lists allow users to create a curated list of twitter accounts the user is interested in. On the list timeline, users can view the stream of tweets from the accounts in that particular list. Users can create their own lists or subscribe to preexisting lists. Listed count describes the number of lists the user is member of.

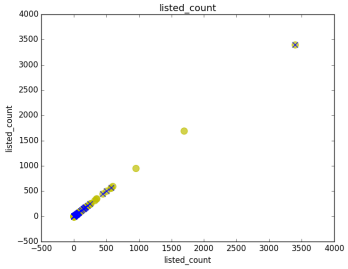

Figure 4.12: Listed Count

There are some outliers to the other extreme of the graph. But overall we see low number of listed count for each group.

4.4.4 *Username Frequency*

Each twitter handle is associated with a user name. For individuals, there are limited number of names that are available so we have these names repeated. A name that is repeated multiple types a strong indicator of the accounts being individuals.

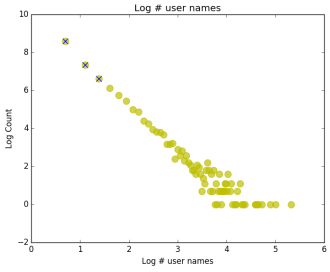

Figure 4.13: Usernames Frequency

User name frequency again not unsurprisingly, follows power law distribution.

4.4.5 HashTag Networks

Hashtags are common terms or phrases that are particular to an event. The users that share common hashtags usually talk about similar phenomenon. These hashtags represnt a concept or an event that occured in a particular region. We make use of these hashtags to build a hashtag network of the users.

We start by collecting all the hashtags by every user. Then for every common hashtag between two users, we add an edge. So the users having multiple hashtags in common have edge weight greater than one. In this way, we create a hashtag network for all the users. Then we calculate the centrality measures and the clustering coefficients on these networks for each user. This is similar to follower or retweet network features. This graph however differs a lot from the other networks described above. For any given hashtag and all the users that mentioned it, we get a complete graph, since all these users used the same hashtag and we have a link between each pair of the users.

4.4.6 Additional Trends

We also found following interesting trends in figure 4.14 that are worth mentioning. User name lengths follows Gaussian like distribution. The second plot is number of user mentions in profile description for each user, third plot is number of ".." in the description. Last one shows the number of exclamations in the user description. These all follow power law distribution.

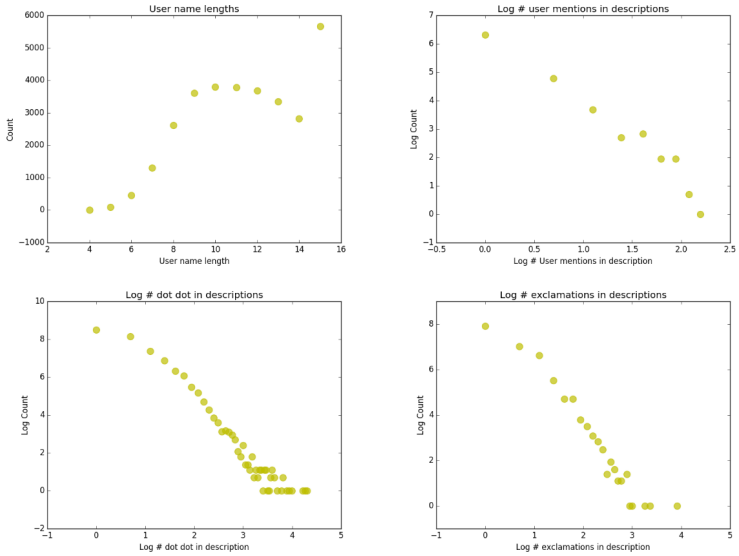

Figure 4.14: Additional Trends

Chapter 5

EXPERIMENTS

In this section the results for detecting groups are presented. We created a set of ground truth labels by manually going through the twitter accounts in the corpus and identifying the political organizations and NGO from the corpus. We also identify celebrities and political leaders which enjoy a large fan base.

5.1 Ground Truth

For labeling the individuals, we created a set of regular expressions that could easily identify the individuals in the corpus. After collecting the ground truth, following data was collected.

Type	Count
Political Organizations	35
NGOs	68
News	62
Individual	30957
Celeb	17
Unlabeled	119534

Table 5.1: Retweet Network Stats

5.2 Classification

We use networkx - Schult and Swart (2008) - for network computation and sci-kit learn - Pedregosa *et al.* (2011) - for training and evaluating classifiers. Following features were used for classification task, see Table 5.2:
We considered only the labelled data for checking the classifier performance. We used the standard precision recall metric to measure the performance of the classifier. We train and test the data using 10 fold cross validation. We observe the cross-validation error in every iteration.

5.3 Evaluation Metric

Since the size of groups is very small, the accuracy of the classifier is always high. Because the classifier tends to predict the population as belonging to the majority class. That is why we consider precision and recall of the classifier. We measure F1 score of the classifier which is harmonic mean of precision and recall.

$$F1score = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

Where precision and recall are given as follows

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

Notice here that the F1 measure gives a good measure to compare the performance here. Precision and recall generally follow inverse relationship. As the precision increases, recall decreases and so on. We want to find the golden middle where F1 measure is maximum.

Retweet Network Features	Retweet Degree Centrality
	Retweet Pagerank Centrality
	Retweet KCore Centrality
	Retweet Clustering Coefficient
Followers network features	Followers Degree Centrality
	Followers Pagerank Centrality
	Followers KCore Centrality
	Followers Clustering Coefficient
User mentions network features	User mentions Degree Centrality
	User mentions Pagerank Centrality
	User mentions KCore Centrality
	User mention Clustering Coefficient
Spatial and Temporal Features	Location entropy
	Location variance
	Timestamp entropy
	Timestamp variance
User behavioral features	Number of lists of user
	Number of favorites by user
	Friends to Followers ratio
	Username frequency of the user
	Hashtag network features

Table 5.2: Features List

5.4 Preprocessing the Data

Since the method proposed here is aimed to be at general classification, we normalize the features to make them scale invariant. For the centrality measures and clustering coefficient, the values are already normalized to be between 0 to 1 except for kcore centrality. The user profile based features such as friends to followers ratio, favorites count, name frequency needs to be normalized. By scaling down the features into [0,1] we preserve the original distribution of the data. Here we used fixed point notation and default precision after the decimal is 26 digits.

5.5 Classifier Performance

We ran following classifiers and we calculated the precision recall on random 10% split of the data. We do a binary classification of the data such the classifier discriminated between groups vs all. Following table 5.3 shows the classifier comparisons arranged in ascending order of F1 score.

Multilayer perceptron has the highest F1 score followed by the random forest. Adaboost and random forest perform almost similar. Precision for random forest seems to be unusually high though and recall is very low. Linear models such as logistic regression are at the bottom of the F1 scores ranking. This might be due to the nature of features. It seems that the linear combination of features is not able to clearly discriminate between organizations and others. That is why we see complex non-linear models such as MLP ad Random forest dominate the classifiers.
As expected the accuracy is high for every classifier since the distribution of classes is imbalanced. Predicting each sample as non-group would give a high accuracy.

Classifier	Precision	Recall	F1 score
SGD	0.490	0.472	0.481
Logistic regression	0.491	0.497	0.494
SVM	0.491	0.500	0.495
Multinomial NB	0.555	0.572	0.562
Decision tree	0.565	0.648	0.587
Passive aggressive	0.659	0.580	0.606
Adaboost	0.742	0.582	0.620
Random forest	0.992	0.583	0.639
MLP	0.795	0.747	0.769

Table 5.3: Classifier Results for Orgs vs Others

We also do a similar analysis of other classes to verify that the features proposed here are able to separate these classes.

Table 5.4 shows results for individuals vs other. Here logistic regression wins over complex model such as MLP. It shows the individuals are linearly separable based on their behavioral features. Which makes a lot of sense because the followship, friends count, tweeting pattern etc is normally seen to be different from other classes.
Since the followers network is expensive to get, we run the classifier without followers features and report the accuracies.

Table 5.6 shows classifier performance for celebrities vs others. Here the decision tree and adaboost perform extremely well. In fact they overfit the model as precision, recall and F1 score is 1. The MLP classifier has very high precision but the recall is

Classifier	Precision	Recall	F1 score
Multinomial NB	0.486	0.500	0.493
Adaboost	0.988	0.556	0.594
SVM	0.989	0.611	0.676
Decision tree	0.641	0.759	0.679
Random forest	0.791	0.663	0.708
SGD	0.892	0.721	0.781
Passive aggressive	0.795	0.827	0.810
MLP	0.870	0.830	0.849
Logistic regression	0.924	0.832	0.872

Table 5.4: Classifier Results for Individuals vs Others

Classifier	Precision	Recall	F1 score
Multinomial NB	0.486	0.500	0.493
SVM	0.988	0.556	0.594
Random forest	0.989	0.611	0.676
Decision tree	0.626	0.857	0.676
Passive aggressive	0.695	0.819	0.740
MLP	0.771	0.771	0.771
Adaboost	0.892	0.721	0.781
SGD	0.992	0.722	0.804
Logistic regression	0.851	0.775	0.808

Table 5.5: Classifier Results for Celebs vs Others - without Followers Features

low. Passive aggressive classifier does little better than MLP and it does not overfit too.

Classifier	Precision	Recall	F1 score
Multinomial NB	0.497	0.500	0.498
SVM	0.497	0.500	0.498
Random forest	0.497	0.500	0.498
SGD	0.750	0.997	0.832
Logistic regression	0.998	0.750	0.833
MLP	0.998	0.750	0.833
Passive aggressive	0.833	0.998	0.899
Decision tree	1.000	1.000	1.000
Adaboost	1.000	1.000	1.000

Table 5.6: Classifier Results for Celebs vs Others

Table 5.7 tries to separate news channels and newspaper accounts from others. Here they are seen to be not distinguishable. This might be due to their skewed characteristics like high follower to followee ratio but unlike celebs they tweet a lot. It becomes difficult for the classifier to discriminate between news and others.

Classifier	Precision	Recall	F1 score
Multinomial NB	0.497	0.500	0.498
Decision tree	0.497	0.494	0.495
SVM	0.497	0.500	0.498
Logistic regression	0.497	0.498	0.498
MLP	0.497	0.500	0.498
Passive aggressive	0.497	0.498	0.498
Random forest	0.497	0.500	0.498
Adaboost	0.497	0.500	0.498
SGD	0.497	0.500	0.498

Table 5.7: Classifier Results for News vs Others

Figure 5.1 shows the wights learned by logistic regression for the features.

(1) Positive Discriminating Features:

Retweet in-degree centrality- People who are retweeted more are likely to be non-individuals

Listed count - On average non-individuals have higher lists count

Retweet pagerank - Non-individuals are retweeted by influential users

Hashtag kcore - In the hashtag network, influential users are at the core of the network

Followers kcore - Non-individuals are in the core of followers network

Followers clustering coefficient - For non-individuals, the followers have more number of friends of friends.

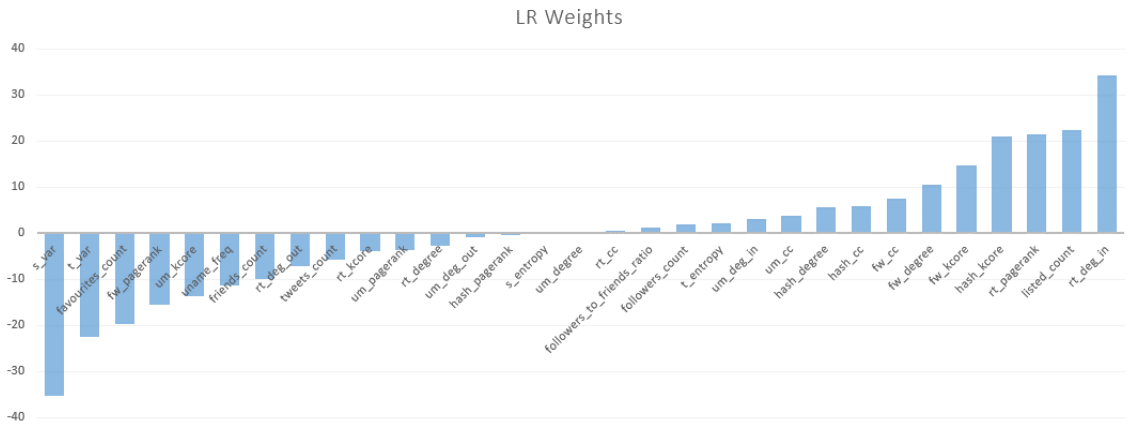

Figure 5.1: LR Weights: Individuals vs All

(2) Negative Discriminating Features:

Spatial variance - Individuals likely to tweet from various locations

Timestamp variance - Individuals have random pattern of tweeting, others are more structured

Favorites count - Individuals have higher number of tweet likes

Followers pagerank - This result is surprising. The followers have high pagerank for individuals. One possible explanation is that a lot the celebrities are counted as individuals in the ground truth labeling.

User mentions kcore - Individuals are more likely to mention each other

(3) Non-discriminating Features:

Hashtag pagerank, spatial entropy, user mentions degree and retweet clustering coefficient are not able to discriminate between the classes.

Chapter 6

CONCLUSION AND FUTURE WORK

This is a first of its kind work which involves detecting political groups and NGOs from Twitter corpus. We showed that we can successfully discriminate between groups vs non-groups without using any language features. The feature set proposed here is applicable to any kind of similar classification or unsupervised learning task. For every node we calculate a new measure of centrality which is comprised of other graph centralities and clustering coefficients. By tuning these centralities, we can detect any kind of node classes. Each of the centrality measures has its own advantages and disadvantages. We can weigh these to distinguish between various kinds of nodes.

The proposed method can be used collect a set of nodes of interest from a graph starting from seed nodes and labelling the predicted nodes on the way. This semi-supervised approach can be beneficial to social analysts. These can be employed by marketing agencies to find the target audience for a particular kind of products based on the existing users that they have. The method can further be improved by getting rid of the noise from the data such as Bots. It would also help the accuracy if we can have completely labelled data. If the data is completely labelled, we can find more patterns in the data which can be used for discriminate between group vs non-groups. Handling the network data is resource intensive task and calculating so many parameters on these humongous networks certainly takes quite a lot amount of time. It can be further speed up by using distributed frameworks such as GraphX which allow the centralities to be calculated in parallel. GPU acceleration methods available out there can also be used to parallelize the degree calculation process.

Recently there has been a lot of work done on convolutional neural networks and

graphs matrices. Using these convonets, we can find similar structures in the graph. Traditionally there have been graph kernels which can measure the similarity between two graphs. We can use such similarity measures to create similarity measures for the graph around the node. This ego graph of a node gives similarity matrix for each node which can be used to cluster the nodes using spectral clustering.

REFERENCES

Barabási, A.-L. and R. Albert, "Emergence of scaling in random networks", science **286**, 5439, 509–512 (1999).

Chu, Z., S. Gianvecchio, H. Wang and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?", in "Proceedings of the 26th annual computer security applications conference", pp. 21–30 (ACM, 2010).

Mislove, A., M. Marcon, K. P. Gummadi, P. Druschel and B. Bhattacharjee, "Measurement and analysis of online social networks", in "Proceedings of the 7th ACM SIGCOMM conference on Internet measurement", pp. 29–42 (ACM, 2007).

Page, L., S. Brin, R. Motwani and T. Winograd, "The pagerank citation ranking: Bringing order to the web.", Tech. rep., Stanford InfoLab (1999).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python", Journal of Machine Learning Research **12**, Oct, 2825–2830 (2011).

Rao, D., D. Yarowsky, A. Shreevats and M. Gupta, "Classifying latent user attributes in twitter", in "Proceedings of the 2nd international workshop on Search and mining user-generated contents", pp. 37–44 (ACM, 2010).

Schult, D. A. and P. Swart, "Exploring network structure, dynamics, and function using networkx", in "Proceedings of the 7th Python in Science Conferences (SciPy 2008)", vol. 2008, pp. 11–16 (2008).

Seidman, S. B., "Network structure and minimum degree", Social networks **5**, 3, 269–287 (1983).

Subrahmanian, V., A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini and F. Menczer, "The darpa twitter bot challenge", Computer **49**, 6, 38–46 (2016).

Travers, J. and S. Milgram, "An experimental study of the small world problem", Sociometry pp. 425–443 (1969).

Varol, O., E. Ferrara, C. A. Davis, F. Menczer and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization", arXiv preprint arXiv:1703.03107 (2017).

Watts, D. J. and S. H. Strogatz, "Collective dynamics of small-worldnetworks", nature **393**, 6684, 440–442 (1998).

Wu, S., J. M. Hofman, W. A. Mason and D. J. Watts, "Who says what to whom on twitter", in "Proceedings of the 20th international conference on World wide web", pp. 705–714 (ACM, 2011).

Zuber, M., "A survey of data mining techniques for social network analysis", International Journal of Research in Computer Engineering & Electronics **3**, 6 (2014).

APPENDIX A

LABELED ACCOUNTS

Political	**People**	**News**	**NGO**
HayatBangla	ImranHSarker	barciknews	SOHAY2002
MdJobayerNaek	sajeebwazed	DhakaTribune	ywbdtweets
RahamanArmanur	BegumZiaBd	bdnews24com	SreepurVillage
DawlaIron	KhaledaZia	ProthomAlo	BaSEBangladesh
Raider_Islamic	BegumZiaBd	BDnews	nirapadorgbd
KhilafahBN	sheikhhasina	bdnews24	ISOCbddhaka
ApniJanenKi1	MdShahriarAlam	bbcbangla	NARRIBangladesh
balakotmedia1	saberhc	dw_bengali	ledarsbd
AlAdiyatMedia2	sufifaruq	VOABANGLA	rdrs_bangla
Ansar_Islam_BD	zapalak	banglanews24com	CMESBD
usama_media	snhera	banglanews_eng	dambgd
borhanrn	ProfGhulamAzam	NewAgeBDcom	ECOTAFTF
AnNashatMedia1	MasoodSayedee	Dhakatimes	nctfbd
AnsarAlIslam5	ShahnurBegum	samakaltw	CharityScf
JihadiGroupBD	sajeebwazed	somoytv	youthprojectbe
balakotmedia1	SamiraHimika	TheDailyInqilab	YPDBangladesh
usama_media	MAlamHanif	Banglatech24	BWCCI
abu__khalid1	ShajahanKhanMP	gvbangla	RCYCTG
umar_mukhtar_3	yeafeshosman	AABangladesh	team_engine
abdullah__abir		BDUpdates	outsourcingscbd
hind_aqsa1		ReutersBiz	SCinBD
mohammadrubel03			Durnibar
info_shibir			YouthOfBD
drkarimbd			czmbd
AbdulZabbarBd			icsComilla
masoodsayedee			TriratnaSangha
ditioalo			YSSEGLOBAL
faraejiandolon			aakfoundation
AndolonNewsATV			sardarmehadi
bnpbangladesh			RIGHTBD
ArDuranta			Smilingfaces_bd
istishon			bsphbd
Mukto_Mona			batf_org
FpcpbRedwan			DayemiFdn
bnpUpdates			Campus_Tweet
bnpbangladesh			rrywa
bipss			Cafbd1
BNP4D			IFMSABangladesh
BdFreedomparty			dccs_bd
raihansumon323			OngshoOrg
DrishtyCtg			hsfbd
bbslbd			bdstudytrust
Justicepartybd			BEDS15

shibirctgnorth1			TanjimulUmmah
info_shibir			AlzheimerBD
AIESEC_NSU			NctfKhulna
fkmarufdu			Ahobanbd
projonmoleague			OxfaminBD
UNinBangladesh			CBSDHAKA
Mukto_Mona			UNICEFBD
SPaRCBangladesh			CAREBdesh
FreeAmaanAzmi			Ashtala
BangladeshLife			usembassydhaka
USAID_BD			ProgressBd
sharifbhuiyan89			GurukulOfficial
IPUparliament			SWFOfficialBD
GurukulOfficial			GurukulGIHT
UKinBangladesh			SASEGOfficial
USAIDBangladesh			GurukulMATS
info_shibir			GurukulSMI
basherkella			PramukhGurukul
BJI_Official			
GurukulLRD			
GurukulDPS			
albd1971			