Interactions of the Twitter Elite: Clustering Politicians, Media Outlets, and Celebrities

Cathy Chen
Adviser: Professor Andrea LaPaugh

Abstract

Social media plays an increasingly significant role in the online and offline worlds, and it is therefore important to understand how users interact on social media platforms. The most visible actors on social media capture a disproportionately large amount of attention from the general public. By studying interactions of highly visible actors, we can obtain a better understanding of the social media landscape.

Many have turned their attention towards three such groups of highly visible actors: celebrities, politicians, and media outlets and some have claimed that the lines between these three groups have become increasingly blurred.

In this project, we study the interactions of the politicians, celebrities, and media outlets with the most followers on Twitter and use the networks created by these interactions to understand clusters and characteristics of these actors on social media. Through our analysis, we aim to answer questions related to the amount of separation between these three groups and the degree to which online influence among the highly-visible elite relates to influence among the general public.

1. Intro

Over the past few years, social media has captured an increasingly large amount of public attention. Many have noted the influence of social media in real world events: interactions on social media platforms served as an impetus for the Arab Spring revolutions, a mobilizing force in the 2008 presidential election, and a popular location for interactions related to the 2016 presidential election [6][1][9].

Twitter, a microblogging website with hundreds of millions of active users, serves as one such social media platform. On Twitter, users can explicitly interact with other users through publicly posted statuses called Tweets. Each Tweet may serve as a reply to another user, a retweet of another status, and may contain mentions of other users. Because these Tweets are available to the general public, we can use them to study how different users interact with each other on Twitter. Each user can "follow" any other user, in which case the following user will receive notifications about Tweets posted by the followed user.

Furthermore, recent events such as the 2016 presidential election have highlighted the increasing overlap between the roles of politicians, celebrities, and media outlets. Members of each group have a significant influence on members of the other two groups, and some have suggested that the boundaries between these groups have become increasingly blurred [18].

In this project, we study Twitter interactions between politicians, media outlets, and celebrities. Previous work has established that a very small number of Twitter users receive a vast majority of attention on Twitter [20]. To gain insight into the parts of Twitter that receive the most attention, we investigate interactions between the most highly followed users of each category. Through our analysis, we seek to understand highly visible users as individuals and as groups; how distinct celebrities, media outlets, and politicians are based on who they interact with on Twitter; and how well the attention of highly-visible members in these groups mirrors that of the general Twitter population.

We hope that insight into these questions can lead to a better understanding of a platform that has captured public attention and influenced societal events.

2. Related Work

Previous researchers have used qualitative and quantitative methods to answer related questions.

Some have compared different types of actors on Twitter and in real life. The Pew Research Center qualitatively examined Twitter conversations, using Tweet content (such as text and URLs) to visualize interactions surrounding various topics [16]. Three of the mapped topics correspond directly to the three types of users that this project studies. The study found polarized clusters around political topics; fragmented clusters around well known topics such as celebrities, and broadcast networks with hubs; and spokes around media outlets. These landscapes give background into how the general Twitter population talks about the groups that we study. Other researchers have used qualitative methods to understand how these actors interact in the real world. For instance, Street notes the increasing overlap between celebrities and politicians, pointing to celebrities who use their popularity to influence political discussions and to politicians who gain elements of celebrityhood through their work [18].

Choi relates these two groups – politicians and celebrities – to media outlets. She finds evidence for the Two-Step Flow of Communication model, which claims that highly visible public opinion leaders serve as a bridge between mass media outlets and the general public [2]. Wu et al connects various groups of highly followed Twitter users in a study that is highly relevant to this project. They study mixing between different groups of users by tracking the spread of URLs, finding a large amount of homogeneity in terms of information spread for celebrities, media outlets, blogs, and organizations [20].

Others have also used network analysis to understand Twitter interactions. Many of these analyses focus on Twitter interactions between politicians, finding that different types of interactions form different types of clusters. For instance, Conover finds that retweet networks produce highly segregated clusters while mention networks form one large heterogeneous cluster [3]. While limited to a single category of actors, these analyses show the promise of using interaction networks to understand Twitter users, and indicate that we should analyze networks created from different types of Twitter interactions.

Lastly, we draw upon a large selection of previous work about methods of network analysis. Previous researchers have established a wide range of metrics used to understand various aspects of a network, and these metrics allow us to analyze individuals, groups, and the network as a whole [11]. Furthermore, previous researchers have created a variety of methods to detect communities in networks [14]. One particular community detection method, spectral clustering, is highly relevant to this project [19]. This method uses a similarity matrix to cluster nodes and can be used to cluster nodes in high-dimensional, non-euclidian spaces.

Existing work provides a rich background for this project, but these studies do not directly answer our questions of interest. Some previous researchers have compared different groups by using a qualitative perspective, or by using Tweet content and URLs. Other researchers, who have used explicit interactions, have largely focused on interactions between members of a specific actor category (usually politicians). My project aims to bridge this divide by looking at explicit interactions between different types of actors. By looking at explicit interactions rather than URL spread, we can determine the specific individuals who initiate and receive interactions, which allows us to group specific users according to their interactions. Furthermore, the ability to look at individual actors allows us to compare actor influence, cluster assignments, and other individual characteristics. By looking at multiple groups, we can understand how highly-visible actors interact in relation to a broader array of societal categories.

3. Approach

We begin by creating networks based on social media interactions between highly visible celebrities, media outlets, and politicians.

We create separate graphs for each type of interaction: mentions, retweets, and replies. Because mentions and replies indicate that the source of the interaction is engaging with the target of the interaction, we use these two interactions as a rough proxy for where a user directs their attention. Since a user's followers view their posts, a user's mentions and retweets can also act as a proxy for where the user directs the attention of those who follow them. When an individual retweets a Tweet, they redistribute the Tweet to all of their followers, so we can consider replies as a rough proxy for where a user directs their support.

We then use a variety of metrics to analyze individuals in terms of their position in the network, and we use other metrics to understand each of the three groups in aggregate. We apply spectral clustering to the interaction networks, which shows us the communities that form within the network.

We analyze these communities to see how well social media communities correspond to categories in real life (using the labels of “politician”, “media outlet”, and “celebrity” to indicate a user’s real-world community classification).

4. Data and Methods

Links to the code, data, and user lists used in this project are included in the Appendix 8.

4.1. Data Retrieval

We begin with three lists of Twitter accounts, each containing the 100 politicians, media outlets, or celebrities with the highest number of followers. We build these lists using a social media analytics company’s pre-defined lists of Twitter users. These lists are separated by category and sorted by number of followers [17]. For each account, we store the account name that identifies them on Twitter, their username, and the number of followers they have.

Using the Twitter REST API, we retrieve as many statuses as possible from each of the accounts in our lists. The Twitter REST API allows us to collect a given user’s statuses in reverse chronological order, up to a maximum of around 3200 statuses [8]. We find any mentions, replies, or retweets in each status. We store the parsed interactions, along with the source and target account names, in a database. We collected these statuses on November 4th, 2016.

4.2. Graph Creation

From the parsed interactions we create three directed graphs: one each for mentions, retweets, and replies. In our graphs, nodes represent the actors whose statuses we retrieved and edges represent interactions parsed from the statuses. In each graph, a directed edge represents an interaction originating from the source node and going towards the target node. For instance, if the account “POTUS” mentions the account “nytimes”, then our mentions graph contains an edge from “POTUS” to “nytimes”.

In our graphs we only include edges between the 300 nodes corresponding to the users on the lists we started with. This allows us to analyze the incoming and outgoing edges from each node in our graphs, which are subgraphs of the Twitter graph induced by the nodes in our list. We focus on the Twitter graph induced by the nodes in our list, rather than attempting to retrieve all possible statuses from everyone that each user in our original list interacted with, to maintain clarity in our graph visualizations and for computational feasibility [1]. We create our graphs using the Python library NetworkX, which allows us to visualize the graphs and to compute a number of metrics [5].

4.3. Network Analysis

We calculate each node's PageRank ranking, betweenness centrality, and closeness centrality. We compute the Pearson correlation coefficient between actor rankings within the network of 300 actors with their ranking in terms of number of followers. We compute this correlation for PageRank ranking, betweenness centrality ranking, and closeness centrality ranking. We compute each node's PageRank ranking and centralities using only interactions with the set of 300 highly-visible actors, while the number of followers involves the general Twitter population. Therefore we use these correlations to understand how a node's influence and centrality among a network of highly visible Twitter actors relates to their influence among the Twitter population in general.

We also analyze the three groups in aggregate by looking at the degree of mixing between different types of nodes. This analysis allows us to understand which categories of actors tend to interact with each other, and whether this differs across different types of interactions.

4.3.1. Visualization We use NetworkX with the Python matplotlib library to visualize each interaction network [7]. We draw each network with a node positioning that minimizes the number of crossing edges, and we use the visualization to gain intuition into the relative positions of nodes from the three categories.

[1] In the statuses we retrieved, the 300 original users mentioned 113004, retweeted 43410, and replied to 38956 distinct users not in our original list. We can retrieve up to around 3200 statuses from each account, and from our list of 300 users we retrieved an average of 2949.24 statuses per person. Twitter rate limits allow us to retrieve a maximum of 200 statuses per request, and to send 900 requests per 15-minute time window [8]. Even if we assume that all users retweeted or replied by users in our list are included in the ones mentioned by users in our list, we would need to retrieve statuses from 113004 other users. This would take a minimum of around $\frac{113004 users \times 2949.24 \frac{statuses}{user}}{200 \frac{statuses}{request} \times 900 \frac{requests}{window} \times 4 \frac{windows}{hour} \times 24 \frac{hours}{day}} = 19.3$ days.

4.3.2. PageRank We use PageRank rankings to understand each actor's influence relative to other actors in the graph. The PageRank algorithm, originally used to evaluate the importance of a web page, ranks the importance of a graph's nodes based on the network's connections. The page rankings correspond to the probability distribution that a "random surfer" (who traverses the graph of webpages by randomly following a link from each webpage it visits, but sometimes decides to jump to a random webpage instead of choosing a link from the current page to follow) visits each page. The rankings satisfy the following equation, where $R(u)$ is the rank of node u, B_u is the set of nodes that point to u, N_v is the total number of edges originating from v, $E(u)$ is a vector over all the nodes that corresponds to the distribution of pages that the random surfer occassionaly jumps to, and c is used for normalization: [12].

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} + cE(u) \tag{1}$$

Others have used the PageRank algorithm to determine the influence of actors on Twitter, substituting interactions for web links and actors for webpages. They found that PageRank rankings correspond to number of followers for the general Twitter population [10].

4.3.3. Betweenness Centrality A node's betweenness centrality tells us the degree to which a node acts as a "bridge" between nodes. We compute this measure using the following formula, where $\sigma(s,t)$ is the number of shortest paths between nodes s and t and $\sigma(s,t|v)$ is the number of shortest paths between nodes s and t that pass through node v [5]:

$$c_B(v) = \Sigma_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{2}$$

4.3.4. Closeness Centrality A node's closeness centrality tells us how closely connected it is to other nodes in the graph, and gives the inverse of the average shortest path to a node. We compute this measure using the following formula, where n equals the number of nodes in the graph and

$d(u,v)$ equals the shortest path between nodes u and v [5]:

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(u,v)}$$

4.3.5. Network Mixing To understand the overall amount of mixing between different types of nodes, we compute the attribute assortativity coefficient. This coefficient measures the degree of connection between different types of nodes in a graph and we compute it for each of our three graphs. The attribute assortativity coefficient r of a graph is given by the following computation, where $\mathbf{e}$ is a matrix in which element e_{ij} is the fraction of edges between vertices of types i and j, $||\mathbf{e}^2||$ is the sum of the elements in the matrix produced by squaring the matrix $\mathbf{e}$, and $Tr\mathbf{e}$ is computed by summing the elements along the diagonal of $\mathbf{e}$ [11]:

$$r = \frac{Tr\mathbf{e} - ||\mathbf{e}^2||}{1 - ||\mathbf{e}^2||} \tag{3}$$

We also look at mixing between different types of nodes more specifically. For each node category, we compute the percentage of interactions originating from members of the category that go towards members of each type of node, and we perform this computation for each type of interaction.

4.4. Clustering

We use scikit learn's spectral clustering algorithm to detect communities in our graphs. As it is used in this project, spectral clustering algorithm receives a symmetric affinity matrix as input, computes a low-dimensional embedding of the affinity matrix, and then performs k-means clustering on the low-dimensional embedding to determine the clusters. [19].

4.4.1. Affinity Matrix Construction To cluster the nodes in a graph, we can input the graph's adjacency matrix as the affinity matrix in the spectral clustering algorithm. However, we represent interaction networks with directed graphs, which have non-symmetric adjacency matrices. Because spectral clustering requires a symmetric affinity matrix, we create an affinity matrix by performing $\mathbf{A} + \mathbf{A}^\top$ symmetrization, where $\mathbf{A}$ is the adjacency matrix of the interaction network in which A_{ij} is

1 if user i interacts with user j and A_{ij} is 0 otherwise. We describe our process for choosing this method of symmetrization in the following subsection 4.4.1.

Choice of Affinity Matrix Construction Method Previous researchers have proposed various methods of symmetrizing adjacency matrices of directed graphs for spectral clustering, and we performed spectral clustering using affinity matrices constructed through pre-existing methods of symmetrization [15].

Symmetrization Methods Tested We tested each symmetrization method on two types of adjacency matrices: the adjacency matrix of the unweighted interaction graph, in which A_{ij} is 1 if user i is the source of an interaction directed towards user j and A_{ij} is 0 otherwise, and the adjacency matrix of the weighted interaction graph, in which A_{ij} is the number of times user i is the source of an interaction directed towards user j. For example, if the account "POTUS" mentioned the account "nytimes" 10 times, and the index of "POTUS" were 0 and the index "nytimes" were 1 in the adjacency matrix, then the unweighted adjacency matrix would have $A_{01} = 1$ but the weighted adjacency matrix would have $A_{01} = 10$.

We tested the following symmetrization methods, and in each of the following descriptions $\mathbf{A}$ is the adjacency matrix of a directed interaction graph:

- **$\mathbf{A}+\mathbf{A}^\top$ symmetrization:** We construct the affinity matrix by adding the affinity matrix to its transpose. This uses information about both incoming and outgoing edges but does not distinguish between them.
- **Bibliometric symmetrization:** We construct the affinity matrix by computing $\mathbf{A}\mathbf{A}^\top + \mathbf{A}^\top\mathbf{A}$, where $\mathbf{A}\mathbf{A}^\top$ captures information about shared outgoing interactions between nodes and $\mathbf{A}^\top\mathbf{A}$ captures information about shared incoming interactions between nodes.
- **Degree-discounted symmetrization:** We construct the affinity matrix by computing $\mathbf{B}+\mathbf{C}$, where $\mathbf{B} = \mathbf{D}_{out}^{-\alpha}\mathbf{A}\mathbf{D}_{in}^{-\beta}\mathbf{A}^\top\mathbf{D}_{out}^{-\alpha}$, $\mathbf{C} = \mathbf{D}_{in}^{-\beta}\mathbf{A}^\top\mathbf{D}_{out}^{-\alpha}\mathbf{A}\mathbf{D}_{in}^{-\beta}$, α and β are discounting parameters which we set to 0.5 based on previous research [15], and D_{in} and D_{out} are diagonal matrices in which the value at index (i,i) respectively indicates the in- or out-degree of node i. As in Bibliographic symmetrization, B and C respectively capture information about shared

outgoing and incoming edges between nodes. However, this method weights the importance of shared interactions according to how rare the interaction is. For instance, if two accounts mention the same node, then the two accounts are considered more similar if the mentioned node is more rarely mentioned.

For each method of constructing the affinity matrix, we computed the average conductance according to the method detailed in the following section 4.4.1. The $\mathbf{A} + \mathbf{A}^\top$ method using the unweighted adjacency matrix of the interaction graph consistently produced the lowest or near-lowest average cluster conductance, which is why we chose this method for the affinity matrix used in this project. Specific average conductance scores are included in the Appendix 8.

Conductance Definition Conductance measures how accurately clusters capture the separation between nodes. For each cluster, we define the conductance score as follows, where C_{in} is the set of nodes inside the cluster and C_{out} is the set of nodes outside of the cluster: $conductance(C_{in}) = \frac{cutcost(C_{in})}{min(volume(C\_in), volume(C\_out))}$. We compute $cutcost(C_{in})$ as the sum of the edges between C_{in} and C_{out}. We compute $volume(C)$ as the sum of the outdegrees of the nodes in C [4].

We define the average cluster conductance as the weighted sum computed by weighting the conductance of each cluster according to the number of users in that cluster.

4.4.2. Cluster Analysis Using this affinity matrix, we perform clustering with 2, 3, and 4 clusters.

We perform clustering with 3 clusters because we begin with three “ground-truth” groups. We also perform clustering with 2 or 4 clusters because this allows us to observe groupings that emerge when the clusters are forced to separate or to join together.

For each cluster, we compute the percentage of the cluster’s nodes that are members of each of the three node categories. These percentages show us the categories of users that tend to be part of the same communities, whether these groupings differ across different types of interactions, and how well the clusters correspond to real world societal categories.

We also use cluster conductance to measure how well the clusters capture communities, homogeneity score to evaluate how well the clusters separate members of different categories, completeness score to evaluate how well the clusters group together members of the same category, and v-measure for an evaluation that balances homogeneity and completeness. More discussion of these evaluation metrics follow in the Results section.

Lastly, we look at the individual users in each cluster. In particular, we look for users that are placed in a cluster that predominantly consists of users of a different type. Using this information, we attempt to anecdotally glean information about the clusters and individuals who are grouped into the "wrong" cluster.

5. Results

The specific metrics resulting from our computations, cluster assignments, and code and data used to produce these results are provided in the Appendix 8.

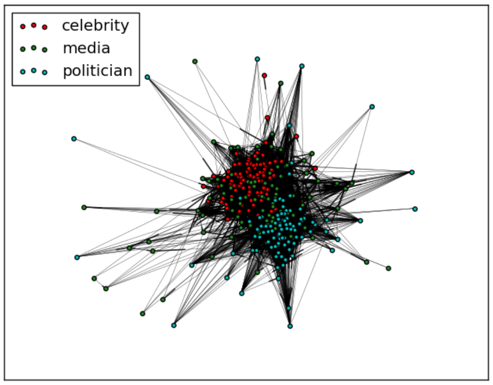
5.1. Graph Visualization

To gain an initial understanding of the interaction networks, we visualize the graph created by the networks for each type of interaction. In these visualizations we distinguish different categories of users by color.

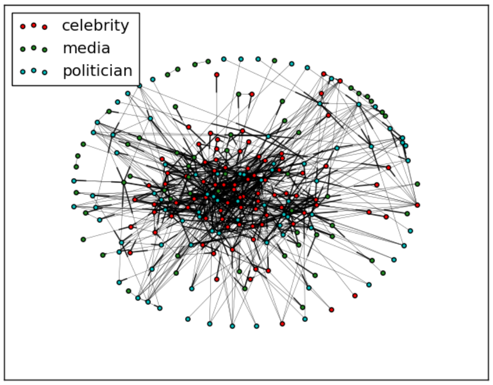
These visualizations, with nodes positioned to minimize the number of crossing edges, show some distinct visual clusters. The network of mentions 1a and of retweets 1c indicate that a cluster of media outlets might serve as a bridge between clusters of celebrities and politicians, while the much sparser network of replies 1b does not show the same strong visual clusters. These results indicate that community detection algorithms may detect clusters that are strongly segregated by user category and that the degree of segregation might differ across different types of interactions.

5.2. Network Positions

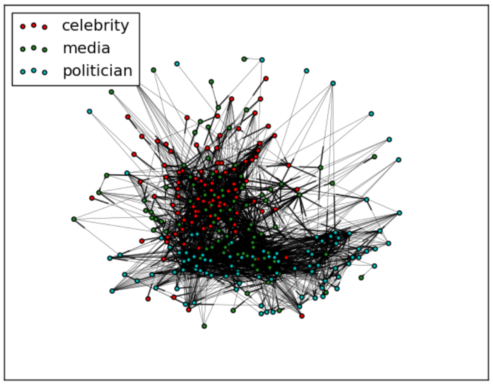
To understand how well influence and centrality within the network of highly-visible users compares to influence in the broader Twitter population, we compute each user's PageRank ranking, betweenness centrality, and closeness centrality. We computed these metrics using edges in our three network interaction graphs and we order the users in descending order according to each of these metrics. Using this ordering, we obtain a rank (between 1 and 300) for each user for each of



(a) Mentions



(b) Replies



(c) Retweets

Figure 1: Graph Visualizations for Interaction Networks.

the three metrics. We also order each user according to the number of followers they have, which gives us another rank (also between 1 and 300) for each user. We compute the Pearson correlation coefficient between the ranks computed according to PageRank, betweenness centrality, closeness centrality, and number of followers. To find the correlation coefficient between a pair of rankings we perform the following computation, where $\mathbf{x}$ contains the rankings given by the first set of rankings, $\mathbf{y}$ contains the rankings given by the second set of rankings, N is the number of users, and $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the average rankings in the respective sets of rankings:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{\mathbf{x}})^2 \sum_{i=1}^{N}(y_i - \bar{\mathbf{y}})^2}} \tag{4}$$

The correlation coefficient ranges between -1 and 1, where a value of 0 means that the two variables are not linearly correlated, a value of 1 means that the two variables are perfectly positively linearly correlated, and a value of -1 means that the two variables are perfectly negatively linearly correlated.

	Betweenness	**Closeness**	**PageRank**	**Followers**
Betweenness	1			
Closeness	0.69	1		
PageRank	0.66	0.25	1	
Followers	0.05	0.03	0.12	1

Table 1: Ranking Correlations for Mentions.

	Betweenness	**Closeness**	**PageRank**	**Followers**
Betweenness	1			
Closeness	0.72	1		
PageRank	0.57	0.23	1	
Followers	0.02	0.08	0.01	1

Table 2: Ranking Correlations for Replies.

We find almost no correlation between users' rankings within the network of highly visible actors, and their rankings (in terms of number of followers) among the general Twitter population 1, 2, 3.

	Betweenness	Closeness	PageRank	Followers
Betweenness	1			
Closeness	0.69	1		
PageRank	0.46	0.03	1	
Followers	-0.05	-0.07	0.01	1

Table 3: Ranking Correlations for Retweets.

	Betweenness	Closeness	PageRank	Followers
Celebrities	147.08	138.06	141.15	56.05
Media	150.93	165.05	148.58	156.1
Politicians	151.96	146.77	160.18	239.35

Table 4: Average Rankings for Mentions by User Category.

We also do not find significant differences between the average rankings of different user categories in our networks, but we find that among the general Twitter population celebrities have the highest average number of followers and politicians the least 4, 5, 6.

5.3. Mixing

To understand how often users from different categories interact each other, we compute the attribute assortativity coefficient for each network.

A high attribute assortativity coefficient of indicates that users associate mostly with other users of the same type, while a low attribute assortativity coefficient indicates that users associate with users of different types. The attribute assortativity coefficients for our networks show that, on average, users reply most often to other users in their category, while they show a slightly greater

	Betweenness	Closeness	PageRank	Followers
Celebrities	119.01	109.77	121.45	56.05
Media	122.81	129.65	120.15	156.1
Politicians	124.11	127.05	124.05	239.35

Table 5: Average Rankings for Replies by User Category.

	Betweenness	Closeness	PageRank	Followers
Celebrities	147.67	150.40	142.39	56.05
Media	147.73	151.06	146.89	156.1
Politicians	139.82	134.01	145.78	239.35

Table 6: Average Rankings for Retweets by User Category.

Interaction Type	Attribute Assortativity Coefficient
Mentions	0.28
Replies	0.55
Retweets	0.32

Table 7: Attribute Assortativity Coefficients for Different Interaction Types.

diversity in their mention and retweet interactions 7.

We further investigated the amount of interaction between different categories of actors by computing the percentage of each category's interactions that were targeted towards users of a given category.

Our results 2 show a high degree of separation between politicians and celebrities across all interaction types, with both groups interacting mostly with other users of the same type, and with media outlets to a lesser extent. For media outlets, interactions have more diversity. With the exception of retweets, for which media outlets also show very homogeneous interactions, media outlets have much more heterogeneity in terms of the types of users they interact with on Twitter.

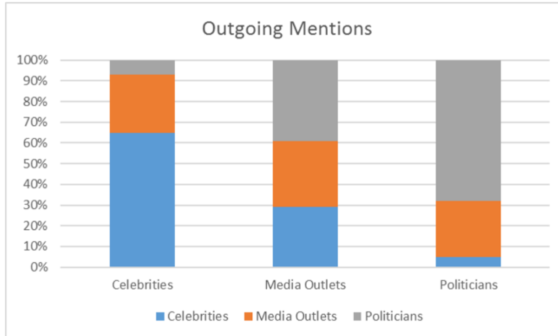
5.4. Clustering

We perform spectral clustering to detect communities among the interaction networks.

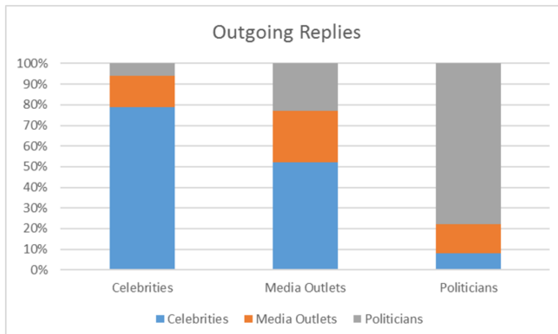
5.4.1. Node Categories In each cluster, we look at the percentage representation of each node group to understand the homogeneity of the communities that emerge in our networks.

When we cluster users into three groups, we find that the resulting clusters differ slightly from the labeled groups of politicians, celebrities, and media outlets 4. We find that spectral clustering produces a cluster that strongly represents politicians and one that strongly represents celebrities, but that the third cluster does not strongly represent media outlets. Rather, media outlets tend to be spread out across all three clusters. The third cluster consists of either a nearly even mix of the three categories (in the case of the network of replies 4b and retweets 4c) or another cluster that strongly represents politicians (in the case of the network of mentions 4a). Based on manual inspection of the users in each category, we did not find any pattern among the users sorted into the clusters with more even representations of all three categories.

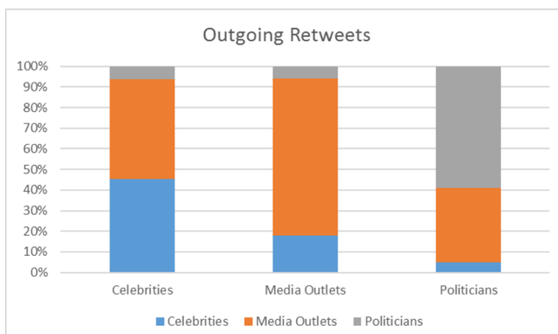
When we cluster users into two 3 or four 5 groups, we see a similar pattern. With the exception



(a) Mentions



(b) Replies



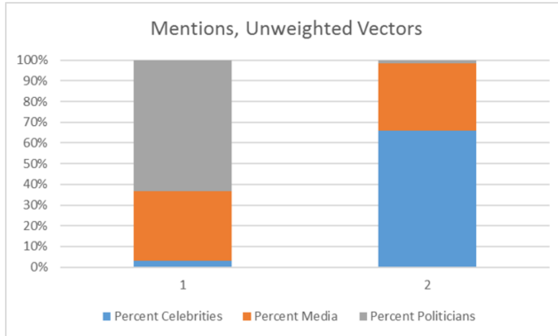
(c) Retweets

Figure 2: Mixing Percentages for Interactions.

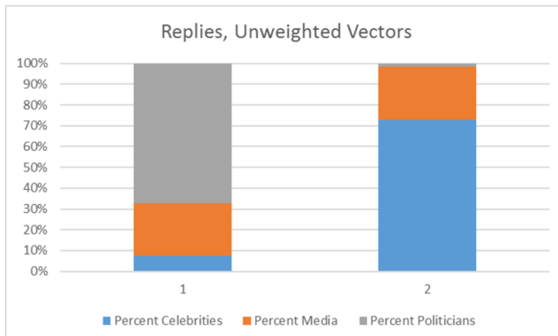
of a single cluster (one of the clusters produced by grouping nodes in the mentions network into four clusters 5a), we generally see that clusters strongly represent either politicians or celebrities, with media outlets split among the various clusters.

Based on manual inspection of individual nodes in each cluster, we were unable to anecdotally determine additional insights about the clusters that users were grouped into.

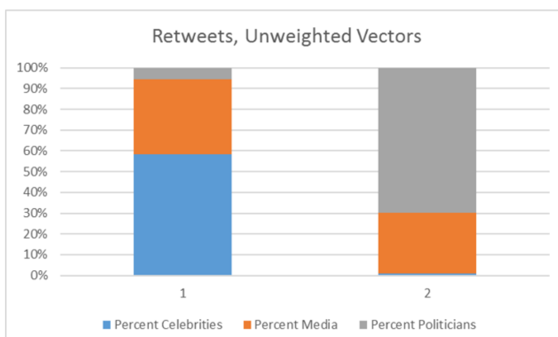
5.4.2. Evaluation To measure the success of our clustering, we look at conductance, homogeneity score, completeness score, and v-measure score. We tested other methods of constructing affinity



(a) Mentions



(b) Replies

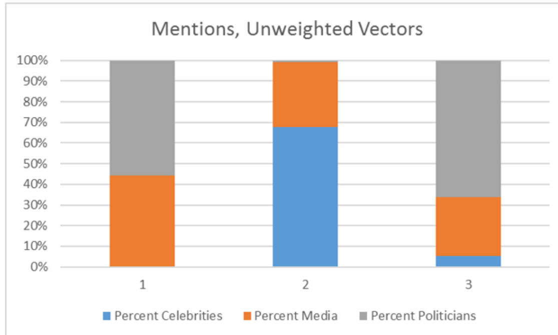


(c) Retweets

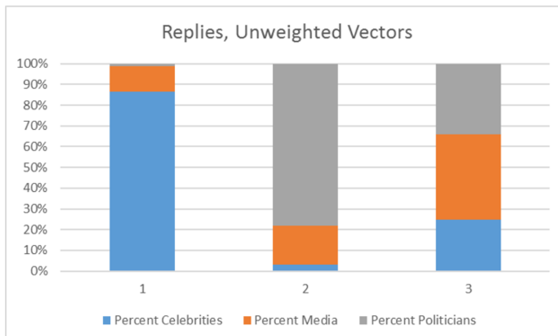
Figure 3: Category Representations for 2 Clusters.

matrices, which are detailed in the "Data and Methods" section 4.4.1. Scores produced by the various affinity matrices we tested are included in the Appendix 8.

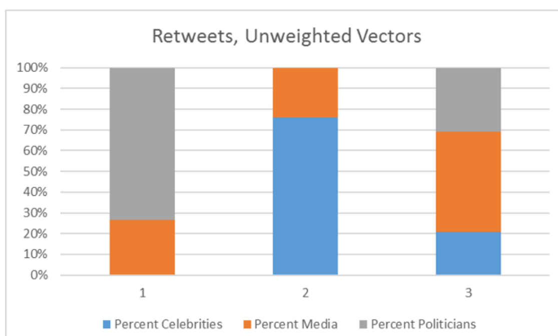
Conductance To evaluate the success of clustering, we compute the conductance score for each set of clusters according to the definition provided in the "Conductance Definition" section 4.4.1. Cluster conductance ranges between 0 to 1, with a conductance of 0 indicating that the clusters have no crossing edges and a conductance of 1 indicating that all the outgoing edges from nodes in one cluster go to nodes in the other cluster. The conductances of clusters produced from our chosen



(a) Mentions



(b) Replies



(c) Retweets

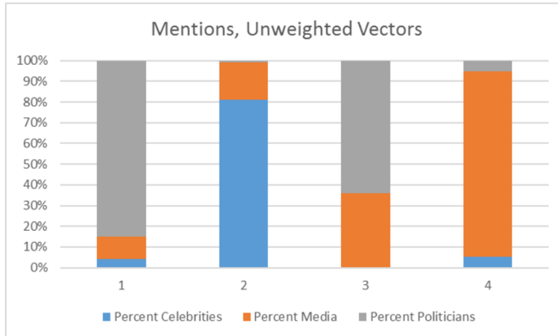
Figure 4: Category Representations for 3 Clusters.

affinity matrix show that the clusters capture the separation between nodes well.

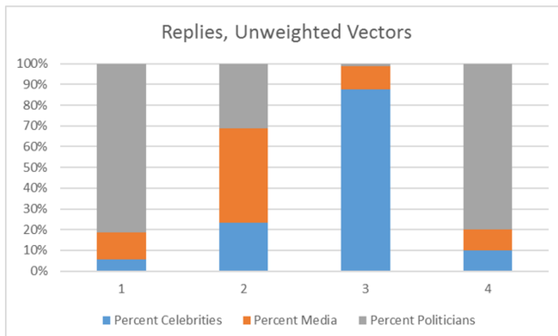
Cluster	**Conductance**
0	0.206
1	0.085
2	0.539
Weighted Average	0.295

Table 8: Conductance for 3 Clusters, Mentions.

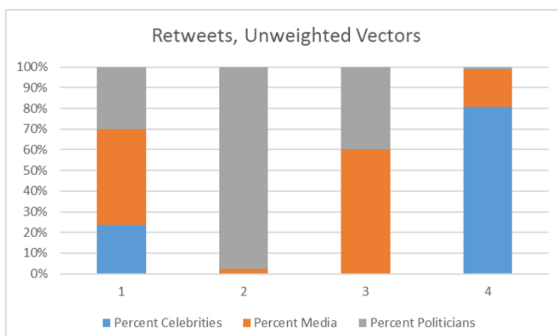
For comparison, we also use the unweighted adjacency matrix of undirected interaction graphs as



(a) Mentions



(b) Replies



(c) Retweets

Figure 5: Category Representations for 4 Clusters.

the affinity matrix in the spectral clustering algorithm. In these affinity matrices, A_{ij} is 1 if nodes i and j interacted and 0 otherwise. This produced average conductance scores of 0.662, 0.608, and

Cluster	Conductance
0	0.150
1	0.055
2	0.67
Weighted Average	0.153

Table 9: Conductance for 3 Clusters, Replies.

Cluster	Conductance
0	0.100
1	0.082
2	0.206
Average	0.114

Table 10: Conductance for 3 Clusters, Retweets.

0.658 for mentions, replies, and retweets respectively. While the results obtained from our chosen clustering method (computing a vector of unweighted incoming and outgoing edges for each user) produces clusters with a non-negligible amount of interaction between the clusters, our clustering method separates users significantly more cleanly than the naive method of clustering based on unweighted adjacency matrix of an undirected interaction graph.

Homogeneity The homogeneity score evaluates how well clusters correspond to categories created by ground truth labels (in this case, we use the original groupings of politicians, media outlets, and celebrities as the ground truth labels). We can find the homogeneity score using the following calculation, where $H(C|K)$ represents the conditional entropy of class labels given the cluster assignments, and $H(C)$ represents the class entropies [13]:

$$h = 1 - \frac{H(C|K)}{H(C)}$$

We can compute $H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} \frac{n_{c,k}}{n} log(\frac{n_{c,k}}{n_k})$ and $H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} log(\frac{n_c}{n})$ where n is the total number of samples, n_k the number of actors in cluster k, n_c the number of actors with ground-truth label c, and $n_{c,k}$ the number of samples from class c assigned to cluster k.

To evaluate how well our clusters capture the three groups based on ground truth labels, we compute the homogeneity score for spectral clustering with three clusters.

Interaction Type	Homogeneity Score
Mentions	0.38
Replies	0.36
Retweets	0.38

Table 11: Homogeneity Scores for 3 Clusters.

These homogeneity scores show that spectral clustering using our chosen method of constructing an affinity matrix captures the ground truth labels we started with reasonably well.

In contrast, the affinity matrix constructed from the unweighted adjacency matrix of an undirected graph representation of the network produces homogeneity scores near zero.

Completeness While the homogeneity score measures the degree to which each cluster only contains members of a single class, the completeness score evaluates the degree to which members of the same class are grouped into the same cluster. We can find the completeness score using the following calculation, where $H(K|C)$ represents the conditional entropy of cluster assignments given class labels, and $H(K)$ represents the cluster entropies [13]:

$$h = 1 - \frac{H(K|C)}{H(K)}$$

We can compute $H(K|C) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|} \frac{n_{c,k}}{n} log(\frac{n_{c,k}}{n_c})$ and $H(K) = -\sum_{k=1}^{|K|} \frac{n_k}{n} log(\frac{n_k}{n})$ where n is the total number of samples, n_k the number of actors in cluster k, n_c the number of actors with ground-truth label c, and $n_{c,k}$ the number of samples from class c assigned to cluster k.

We compute the homogeneity score for spectral clustering with three clusters and similarly find that the clusters generally groups members of the same class into the same cluster 12. For contrast, we provide the completeness scores of clustering using an unweighted adjacency matrix of the undirected network as the affinity matrix. These clusters had completeness scores near zero.

Interaction Type	Completeness Score
Mentions	0.40
Replies	0.41
Retweets	0.34

Table 12: Completeness Scores for 3 Clusters.

V-measure V-measure represents a combination of completeness and homogeneity, and we compute this metric as $v = 2 \times \frac{h \times c}{h+c}$, where h represents the homogeneity score and c represents the completeness score. The V-measure scores of our clusters 13 show that the clusters do a reasonably good job creating clusters that group nodes in the same category together and that produce clusters that contain nodes in the same category. Again, for comparison, the V-measure scores of clusters produced with the unweighted adjacency matrix of the undirected interaction graph are near zero.

Interaction Type	V-measure Score
Mentions	0.39
Replies	0.38
Retweets	0.32

Table 13: V-measure Scores for 3 Clusters.

6. Discussion

6.1. Conclusions

In this project, we study the interactions of highly visible politicians, media outlets, and celebrities on Twitter. Through our analysis, we seek to understand where groups of highly visible actors direct their attention and support towards, and what clusters emerge within these networks. For politicians and celebrities, we find highly homogeneous interactions across all types of interactions. For media outlets, we find homogeneity in the network of retweets but significantly more heterogeneous interactions in the networks of mentions and replies. By using spectral clustering, we find that politicians and celebrities tend to segregate themselves into separate groups, while media outlets do not show the same degree of segregation. We also look at the relationship between importance among the network of highly-visible users to importance among the general Twitter population. We use betweenness centrality, closeness centrality, and PageRank ranking as a measure of importance in the network of interactions between the 300 nodes we started with and we use number of followers as a measure of importance among the general Twitter population. Using these measures, we find virtually no correlation between importance in the network of highly visible users and importance in the general Twitter population.

These findings can be applied towards understanding the actions of specific users on Twitter and to further understand the impact that Twitter has on real world events. We could use these results to sort unlabeled Twitter actors based on their interactions by looking at the user type that is most heavily represented in the cluster they fall into. Furthermore, we can use these results to further our understanding of how the focus of the Twitter community mirrors society as a whole, both online and in the real world.

6.2. Limitations and Possible Future Work

Our data collection method introduces some limitations. In order to maintain a reasonable computation time for computing metrics and visualizing graphs, we analyzed the graph induced by our list of 300 users. Because we did not consider interactions with any other users, it is likely that some other influential actors were not included in our analysis. We believe that the subset we selected constitute a reasonable population for answering our questions of interest – which relate to the interactions of highly visible politicians, media outlets, and celebrities – but extending our analysis by including additional actors on Twitter could reveal additional insights.

Furthermore, the Twitter API limits the number of statuses we can collect per user. The ages of accounts we studied, the frequency of each user's Tweets, and the frequency of each user's interactions with other users differ widely. Therefore the collected statuses cover different time ranges for different accounts, and the number of interactions originating from each user differs significantly. We believe that these differences constitute a feature of these users' Twitter interactions, rather than a bug in the method of analysis, but our results could be extended by normalizing our data by time, interaction frequency, or Tweet quantity.

Lastly, we collected our data shortly before the 2016 presidential election (on November 4th, 2016), which may have had an impact on the results of our analysis. Twitter played a significant role in this election, and many high-profile interactions played out on Twitter in the months surrounding the election. We re-ran our analysis on a set of newly collected data (collected on December 30, 2016). The evaluation metrics for clustering the newly collected data had similar values to those for the originally collected data. We also found similar correlation coefficients between rankings of influence within the network (as determined by PageRank, betweenness centrality, and closeness centrality) and among the general Twitter population (as determined by number of followers). Data and metrics for the newly collected datasets are available with the originally collected data, as described in the Appendix 8. However, given that this data comes between the election and the presidential inauguration, it is possible that our results are still heavily influenced by the presidential election. We believe that it would be informative to re-run our analysis for data collected at a later date. It would be very interesting to run this analysis for data collected at different points along a much longer span of time. This would allow us to analyze trends in groups' interactions. Furthermore, it could show the influence of various real world events, and could potentially be used as a method for using the interactions of highly visible actors on Twitter to detecting influential events in the real world.

7. Acknowledgments

Thank you to Professor Andrea LaPaugh, who advised this project and provided extraordinarily helpful feedback and advice. Also, thank you to the members of the IW03 and SOC596 seminars who provided feedback along the way, and for the Princeton CS department for supporting this project.

References

[1] D. Carr, "How obama tapped into social networks' power," Washington, DC, USA. Available: http://www.nytimes.com/2008/11/10/business/media/10carr.html

[2] S. Choi, "The two-step flow of communication in twitter-based public forums," *Social Science Computer Review*, 2014. Available: http://ssc.sagepub.com/content/early/2014/11/06/0894439314556599.abstract

[3] M. Conover *et al.*, "Political polarization on twitter," in *5th International Conference on Weblogs and Social Media (ICWSM), 2011*, Barcelona, Spain, 07/17-21/2011 2011.

[4] D. F. Gleich, "Hierarchical directed spectral graph partitioning," Information Networks, Stanford University, Final Project, 2005, 2006, cited over 6 times. Available: http://www.cs.purdue.edu/homes/dgleich/publications/Gleich2005-hierarchicaldirectedspectral.pdf

[5] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, Aug. 2008, pp. 11–15.

[6] P. Howard *et al.*, "Opening closed regimes: What was the role of social media during the arab spring?" Project on Information Technology and Political Islam, Tech. Rep., 2011. Available: http://dx.doi.org/10.2139/ssrn.2595096

[7] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[8] T. Inc, "Twitter developer documentation," https://dev.twitter.com/rest/reference/get/statuses/user_timeline.

[9] M. Isaac and S. Ember, "For election day influence, twitter ruled social media," Washington, DC, USA. Available: http://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter-ruled-social-media.html?_r=0

[10] H. Kwak *et al.*, "What is twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600. Available: http://doi.acm.org/10.1145/1772690.1772751
[11] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.
[12] L. Page *et al.*, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. Available: http://ilpubs.stanford.edu:8090/422/
[13] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[14] M. Porter, J.-P. Onnela, and P. Mucha, "Communities in networks," *Notices of the American Mathematical Society*, vol. 56, no. 9, pp. 1082–1097, 1164–1166, 2009.
[15] V. Satuluri and S. Parthasarathy, "Symmetrizations for clustering directed graphs," in *Proceedings of the 14th International Conference on Extending Database Technology*, ser. EDBT/ICDT '11. New York, NY, USA: ACM, 2011, pp. 343–354. Available: http://doi.acm.org/10.1145/1951365.1951407
[16] M. A. Smith *et al.*, "Mapping twitter topic networks: From polarized crowds to community clusters," Pew Research Center, Washington, DC, USA, Tech. Rep., 2014. Available: http://www.pewinternet.org/files/2014/02/PIP_Mapping-Twitter-networks_022014.pdf
[17] S. Solutions, "Twitter statistics directory," Available at https://www.socialbakers.com/statistics/twitter/.
[18] J. Street, "Celebrity politicians: Popular culture and political representation," *The British Journal of Politics and International Relations*, vol. 6, no. 4, pp. 435–452, 2004. Available: http://bpi.sagepub.com/content/6/4/435.abstract
[19] U. von Luxburg, "A tutorial on spectral clustering," 2007.
[20] S. Wu *et al.*, "Who says what to whom on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 705–714. Available: http://doi.acm.org/10.1145/1963405.1963504

8. Appendix

8.1. Implementation: Code and Data

All code and data used to conduct this project, as well as interactive visualizations of the network, are available at `https://github.com/cchen23/Fall-2016-IW`.

8.2. User Information

In this section we provide the names, usernames, and number of followers for each account we retrieved statuses from, separated by category.

Table 14: Celebrities List.

Name	Username	Number of Followers		Name	Username	Number of Followers
50cent	50Cent	8 723 410		ActuallyNPH	Neil Patrick Harris	25 593 162
aliciakeys	Alicia Keys	24 030 712		andersoncooper	Anderson Cooper	8 649 168
aplusk	Ashton Kutcher	17 548 621		ArianaGrande	Ariana Grande	42 217 817
ashleytisdale	Ashley Tisdale	13 252 356		AustinMahone	Austin Mahone	9 622 038
azizansari	Aziz Ansari	10 829 752		Beyonce	Beyoncé	14 604 498
BigSean	Sean Don	9 244 680		BillGates	Bill Gates	31 208 894
blakeshelton	Blake Shelton	18 108 880		britneyspears	Britney Spears	47 938 333

Continued on next page

Table 14 – continued from previous page

Name	Username	Number of Followers		Name	Username	Number of Followers
BrunoMars	Bruno Mars	26 460 279		camerondallas	Cameron Dallas	9 083 089
carmeloanthony	Carmelo Anthony	8 370 386		channingtatum	Channing Tatum	8 580 999
charliesheen	Charlie Sheen	11 668 919		chrisbrown	Chris Brown	17 678 737
ciara	Ciara	7 316 297		ConanOBrien	Conan O'Brien	21 989 171
danieltosh	Daniel Tosh	25 102 683		ddlovato	Demi Lovato	38 874 223
Eminem	Marshall Mathers	20 781 429		EvaLongoria	Eva Longoria Baston	7 420 901
iamdiddy	Sean Diddy Combs	11 719 947		IAMQUEENLATIFAH	Queen Latifah	7 210 828
iamwill	I.Am+	13 622 394		IGGYAZALEA	Iggy Azalea	7 117 671
JColeNC	J. Cole	8 778 803		jessicaalba	Jessica Alba	9 371 980
JessicaSimpson	Jessica Simpson	7 260 030		JimCarrey	Jim Carrey	15 336 262
jimmyfallon	Jimmy Fallon	42 982 092		jimmykimmel	Jimmy Kimmel	8 007 499
JLo	Jennifer Lopez	37 532 371		joejonas	J O E J O N A S	8 319 951
JohnCena	John Cena	8 756 823		johnlegend	John Legend	8 258 195
jtimberlake	Justin Timberlake	56 753 923		justinbieber	Justin Bieber	88 699 317
kanyewest	Kanye West	26 246 721		katyperry	Katy Perry	93 369 751
KDTrey5	Kevin Durant	15 060 923		kelly_clarkson	Kelly Clarkson	11 188 966
KendallJenner	Kendall	19 978 386		kendricklamar	Kendrick Lamar	7 197 693
KevinHart4real	Kevin Hart	30 841 367		khloekardashian	Khloé	22 208 174
KimKardashian	Kim Kardashian West	48 376 187		KingJames	Lebron James	33 244 051
kobebryant	Kobe Bryant	10 767 414		kourtneykardash	Kourtney Kardashian	21 163 315
KrisJenner	Kris Jenner	7 696 281		KylieJenner	Kylie Jenner	18 244 423
ladygaga	Joanne	63 997 129		LeoDiCaprio	Leonardo Dicaprio	16 187 169
LilTunechi	Lil Wayne Weezy F	28 130 664		lindsaylohan	Lindsay Lohan	9 312 088
LMFAO	Lmfao	8 273 806		Ludacris	Ludacris	11 646 194
LukeBryanOnline	Luke Bryan	7 680 293		MarcAnthony	Marc Anthony	8 905 101
MariahCarey	Mariah Carey	16 689 205		maroon5	Maroon 5	13 298 757
MileyCyrus	Miley Ray Cyrus	30 804 092		mindykaling	Mindy Kaling	7 955 002
MirandaCosgrove	Miranda Cosgrove	7 533 379		NICKIMINAJ	Nicki Minaj	20 642 084
nickjonas	Nick Jonas	10 875 012		Oprah	Oprah Winfrey	34 336 209
ParisHilton	Paris Hilton	13 855 030		Pharrell	Pharrell Williams	9 349 444
Pink	P!Nk	28 246 806		pitbull	Pitbull	22 767 655
RobertDowneyJr	Robert Downey Jr	7 554 969		RyanSeacrest	Ryan Seacrest	14 961 338
SarahKSilverman	Sarah Silverman	8 898 502		selenagomez	Selena Gomez	45 945 224
serenawilliams	Serena Williams	7 097 604		SethMacFarlane	Seth Macfarlane	10 946 541
SHAQ	Shaq	12 702 287		ShawnMendes	Shawn Mendes	7 618 905
SnoopDogg	Snoop Dogg	14 636 989		StephenAtHome	Stephen Colbert	11 965 120
StephenCurry30	Stephen Curry	7 191 327		SteveMartinToGo	Steve Martin	7 832 018
taylorswift13	Taylor Swift	81 148 129		TheEllenShow	Ellen Degeneres	62 588 079
TheRock	Dwayne Johnson	10 520 023		tomhanks	Tom Hanks	12 210 302
TreySongz	Trey Songz	10 341 547		Tyga	T-Raww	7065252
tyrabanks	Tyra Banks	13 515 832		Usher	Usher Raymond Iv	11 402 470
VictoriaJustice	Victoria Justice	9 457 260		wizkhalifa	Wiz Khalifa	27 804 845
xtina	Christina Aguilera	16 138 736		ZacEfron	Zac Efron	12 928 200

Table 15: Media List.

Name	Username	Number of Followers		Name	Username	Number of Followers
ABC	Abc News	7 718 914		AP	The Associated Press	8 736 830
BBCBusiness	Bbc Business	1 813 451		BET	Bet	1 905 469

Continued on next page

Table 15 – continued from previous page

Name	Username	Number of Followers		Name	Username	Number of Followers
billboard	Billboard	4 128 761		business	Bloomberg	3 438 014
businessinsider	Business Insider	1 653 786		BuzzFeed	Buzzfeed	3 955 754
BW	Businessweek	1 393 318		CBSNews	Cbs News	4 870 269
CNBC	Cnbc	2 435 953		cnni	Cnn International	5 628 045
CNNMoney	Cnnmoney	1 651 999		CNTraveler	Condé Nast Traveler	1 468 814
ComedyCentral	Comedy Central	1 461 235		Cosmopolitan	Cosmopolitan	1 514 205
digg	Digg	1 530 524		Discovery	Discovery	5 688 007
ELLEmagazine	Elle Magazine	5 992 669		enews	E! News	10 328 085
Entrepreneur	Entrepreneur	2 640 174		ESPNcricinfo	Espncricinfo	3 943 775
ESPNDeportes	Espn Deportes	1 831 155		ESPNStatsInfo	Espn Stats & Info	1 442 252
EsquireNetwork	Esquire Network	1 586 466		EW	Entertainment Weekly	5 752 537
FastCompany	Fast Company	2 159 426		foodandwine	Food & Wine	4 970 351
FortuneMagazine	Fortune	2 144 137		FoxNews	Fox News	11 238 760
FOXSports	Fox Sports	1 721 963		gadgetlab	Wired Gadget Lab	1 940 731
gameinformer	Game Informer	1 860 001		goodhealth	Health	3 197 433
gov	Twitter Government	2 274 522		HarvardBiz	Harvard Biz Review	3 300 734
HBO	Hbo	1 572 360		HISTORY	History	1 550 893
Inc	Inc.	1 893 522		latimes	Los Angeles Times	2 232 928
LIFE	Life	4 562 949		lifehacker	Lifehacker	3 489 963
marieclaire	Marie Claire	2 176 416		MarketWatch	Marketwatch	3 053 908
Marvel	Marvel Entertainment	4 269 365		mashable	Mashable	7 822 855
MensHealthMag	Men'S Health Mag	3 519 089		NatGeoPhotos	Nat Geo Photography	2 391 940
NatGeoTravel	Natgeotravel	3 017 833		NBAonTNT	Nba On Tnt	1 693 067
NBATV	Nba Tv	2 707 785		nbc	Nbc	1 433 728
NBCNews	Nbc News	3 876 084		Newsweek	Newsweek	3 021 471
NewYorker	The New Yorker	6 801 094		nflnetwork	Nfl Network	2 762 418
NickelodeonTV	Nickelodeon	3 978 544		NPR	Npr	6 245 878
nprpolitics	Npr Politics	2 368 651		nytimes	The New York Times	30 936 311
nytimesarts	New York Times Arts	2 256 794		nytimesbooks	New York Times Books	4 031 507
nytimesworld	New York Times World	1 538 066		parenting	Parenting.Com	1 587 085
parentsmagazine	Parents Magazine	3 764 262		PBS	Pbs	2 248 400
people	People Magazine	7 374 868		politico	Politico	2 075 384
RealSimple	Real Simple	1 469 468		RollingStone	Rolling Stone	5 279 722
SAI	Bi Tech	1 383 982		sciam	Scientific American	1 765 868
ScienceChannel	Science Channel	2 053 438		Slate	Slate	1 611 624
TechCrunch	Techcrunch	7 691 459		TeenVogue	Teen Vogue	2 656 688
TheAtlantic	The Atlantic	1 355 557		thehill	The Hill	1 403 577
TheNextWeb	The Next Web ??????	1 694 885		THR	Hollywood Reporter	2 241 497
tmagazine	T Magazine	2 103 141		TMZ	Tmz	4 245 203
travelchannel	Travel Channel	2 310 537		TravelLeisure	Travel + Leisure	2 659 608
twitterapi	Twitter Api	6 120 169		UniNoticias	Univision Noticias	1 467 289
Univision	Univision	2 449 479		USATODAY	Usa Today	2 696 144
usweekly	Us Weekly	2 020 664		VanityFair	Vanity Fair	4 192 283
voguemagazine	Vogue Magazine	11 792 438		washingtonpost	Washington Post	7 589 534
weatherchannel	The Weather Channel	2 141 398		WhoWhatWear	Who What Wear	2 172 499
WIRED	Wired	7 116 434		WIREDScience	Wired Science	1 955 397
WomensHealthMag	Women'S Health	4 232 499		WSJ	Wall Street Journal	11 961 330
wwd	Wwd	2 946 016		younghollywood	Young Hollywood	3 009 372

Table 16: Politicians List.

Name	Username	Number of Followers		Name	Username	Number of Followers
BarackObama	Barack Obama	77 884 582		realDonaldTrump	Donald J. Trump	12 227 110
POTUS	President Obama	10 681 955		HillaryClinton	Hillary Clinton	9 536 475
billclinton	Bill Clinton	6 074 980		BernieSanders	Bernie Sanders	3 615 183
algore	Al Gore	2 987 044		SenSanders	Bernie Sanders	2 611 116
nasahqphoto	Nasa Hq Photo	2 367 358		RealBenCarson	Dr. Ben Carson	2 168 028
SenJohnMcCain	John Mccain	2 041 502		JohnKerry	John Kerry	1 960 269
CoryBooker	Cory Booker	1 811 500		MittRomney	Mitt Romney	1 804 714
newtgingrich	Newt Gingrich	1 673 001		tedcruz	Ted Cruz	1 644 787
marcorubio	Marco Rubio	1 614 541		SpeakerBoehner	John Boehner	1 294 929
SarahPalinUSA	Sarah Palin	1 290 672		JoeBiden	Joe Biden	1 252 999
GavinNewsom	Gavin Newsom	1 243 200		JerryBrownGov	Jerry Brown	1 059 706
SenWarren	Elizabeth Warren	965 964		JoeTrippi	Joe Trippi	960 136
RandPaul	Dr. Rand Paul	907 652		GOP	Gop	868 567
SpeakerRyan	Paul Ryan	867 047		NancyPelosi	Nancy Pelosi	846 510
SenTedCruz	Senator Ted Cruz	803 738		JebBush	Jeb Bush	762 211
elizabethforma	Elizabeth Warren	761 098		GovChristie	Governor Christie	755 308
PRyan	Paul Ryan	751 219		CarlyFiorina	Carly Fiorina	700 707
TheDemocrats	The Democrats	679 121		BilldeBlasio	Bill De Blasio	655 619
KarlRove	Karl Rove	654 349		RonPaul	Ron Paul	596 749
donnabrazile	Donna Brazile	566 124		AmbassadorRice	Susan Rice	563 166
AllenWest	Allen West	556 567		JohnEMichel	John Michel	526 296
TheRevAl	Reverend Al Sharpton	505 722		GovMikeHuckabee	Gov. Mike Huckabee	498 556
Reince	Reince Priebus	492 052		davidaxelrod	David Axelrod	488 077
MufiHannemann	Mufi Hannemann	487 094		LouisFarrakhan	Minister Farrakhan	457 889
JohnKasich	John Kasich	445 162		Astro_Jose	Jose Hernandez	411 424
THEHermanCain	Herman Cain	391 180		Lagarde	Christine Lagarde	370 456
GovGaryJohnson	Gov. Gary Johnson	368 064		SalahiTareq	Tareq Salahi	364 179
SenatorReid	Senator Harry Reid	363 022		HouseDemocrats	House Democrats	359 247
timkaine	Senator Tim Kaine	356 874		DWStweets	D Wasserman Schultz	356 761
GovernorPerry	Rick Perry	344 251		TGowdySC	Trey Gowdy	329 498
RBReich	Robert Reich	300 649		RickSantorum	Rick Santorum	273 803
PreetBharara	Us Attorney Bharara	271 790		alfranken	Al Franken	266 209
MicheleBachmann	Michele Bachmann	262 590		JohnKingatED	John King	262 504
EricCantor	Eric Cantor	260 859		FreedomWorks	Freedomworks	258 619
LindaSuhler	Linda Suhler Ph.D.	253 116		BorisJohnson	Boris Johnson	252 002
SenMikeLee	Mike Lee	244 977		BobbyJindal	Gov. Bobby Jindal	243 624
ScottWalker	Scott Walker	236 624		VoteRocky2016	Roque De La Fuente	233 194
DarrellIssa	Darrell Issa	227 530		JimDeMint	Jim Demint	226 159
GeorgeHWBush	George Bush	209 330		KasimReed	Kasim Reed	203 014
repjohnlewis	John Lewis	199 634		GabbyGiffords	Gabrielle Giffords	198 236
dccc	Dccc	194 212		SenGillibrand	Kirsten Gillibrand	193 501
BishopNoelJones	Bishop Noel Jones	191 493		CondoleezzaRice	Condoleezza Rice	186 300
jeffarazzi	Jeff Johnson	185 027		wendydavis	Wendy Davis	180 256
RepRonPaul	Ron Paul	178 856		TPPatriots	Tea Party Patriots	176 983
MassGovernor	Charlie Baker	173 613		GovWalker	Governor Walker	161 852
SenSchumer	Chuck Schumer	155 727		nikkihaley	Nikki Haley	155 717
TeamCantor	Team Cantor	152 142		jasoninthehouse	Jason Chaffetz	149 903
ChrisChristie	Chris Christie	146 191		ericgarcetti	Eric Garcetti	145 719
McFaul	Michael Mcfaul	142 835		ForAmerica	Foramerica	142 806
billpostmus	Bill Postmus	141 948		Madam_President	Madam President	137 994

8.3. Network Analysis

8.3.1. User Rankings In this section we provide user rankings for each type of interaction based on betweenness centrality, closeness centrality, and PageRank score. We provide these rankings for the first 50 users, ordered alphabetically.

Note: Some users are not included in some interaction networks. This happens if, based on the statuses we collected, a user is not involved in any of the network's type of interactions.

Table 17: Mentions Rankings.

User	PageRank Score	PageRank Ranking	Betweenness Centrality	Betweenness Ranking	Closeness Centrality	Closeness Ranking
50cent	0.002	143	0.002	177	0.426	164
abc	0.004	67	0.011	46	0.408	197
actuallynph	0.001	188	0.006	105	0.46	79
alfranken	0.001	214	0.002	194	0.449	101
algore	0.001	173	0.003	164	0.466	63
aliciakeys	0.002	103	0.006	104	0.464	67
allenwest	0.001	247	0	269	0.314	288
ambassadorrice	0.001	198	0.001	198	0.43	152
andersoncooper	0.002	142	0.017	19	0.468	62
ap	0.003	89	0.01	51	0.404	204
aplusk	0.001	202	0.008	64	0.462	73
arianagrande	0.005	41	0.002	174	0.435	141
ashleytisdale	0.001	237	0.002	180	0.444	116
astro_jose	0.001	291	0	285	0.393	219
austinmahone	0.001	225	0	238	0.397	212
azizansari	0.001	195	0.005	114	0.461	77
barackobama	0.018	7	0.025	11	0.447	106
bbcbusiness	0.001	294	0	292	0.323	282
berniesanders	0.009	19	0.014	26	0.441	127
bet	0.003	96	0.007	80	0.457	83
beyonce	0.004	57	0	278	0	298
bigsean	0.003	84	0.002	176	0.446	109
billboard	0.011	13	0.009	55	0.482	31
billclinton	0.011	12	0.026	7	0.441	125
billdeblasio	0.001	184	0.001	216	0.352	264
billgates	0.002	146	0.008	68	0.439	131
billpostmus	0.001	282	0.002	188	0.488	19
bishopnoeljones	0.001	297	0	297	0.328	278
blakeshelton	0.004	59	0.003	163	0.428	159
bobbyjindal	0.001	226	0.003	148	0.463	72
borisjohnson	0.001	254	0	276	0.306	293
britneyspears	0.003	93	0.004	137	0.45	99
brunomars	0.002	138	0.005	116	0.443	120
business	0.002	104	0.001	223	0.384	229

Continued on next page

Table 17 – continued from previous page

User	PageRank Score	PageRank Ranking	Betweenness Centrality	Betweenness Ranking	Closeness Centrality	Closeness Ranking
businessinsider	0.001	197	0.006	101	0.443	122
buzzfeed	0.003	79	0.007	77	0.418	182
bw	0.001	183	0.003	154	0.428	160
camerondallas	0.001	229	0.001	215	0.368	254
carlyfiorina	0.002	159	0.005	110	0.465	65
carmeloanthony	0.003	75	0.003	146	0.434	143
cbsnews	0.002	158	0.008	63	0.445	112
channingtatum	0.001	180	0.003	162	0.449	100
charliesheen	0.001	219	0.002	187	0.436	139
chrisbrown	0.004	62	0.001	210	0.434	142
chrischristie	0.001	176	0.002	178	0.487	22
ciara	0.001	178	0.003	159	0.454	87
cnbc	0.002	115	0.007	74	0.38	233
cnni	0.001	243	0.001	228	0.388	224
cnnmoney	0.001	217	0.008	67	0.472	53
cntraveler	0.001	271	0	242	0.377	243

Table 18: Replies Rankings.

User	PageRank Score	PageRank Ranking	Betweenness Centrality	Betweenness Ranking	Closeness Centrality	Closeness Ranking
50cent	0.004	71	0.001	108	0.174	97
abc	0.002	159	0	179	0	211
actuallynph	0.002	131	0.002	102	0.204	45
alfranken	0.001	190	0.001	111	0.154	127
algore	0.002	163	0.007	61	0.176	93
aliciakeys	0.003	82	0.005	72	0.225	19
allenwest	0.001	187	0	134	0.004	177
ambassadorrice	0.001	238	0	229	0.141	147
andersoncooper	0.001	225	0	197	0.196	58
ap	0.003	95	0	218	0	227
aplusk	0.002	133	0.006	63	0.213	33
arianagrande	0.014	13	0.01	45	0.216	27
ashleytisdale	0.003	97	0.02	23	0.214	30
astro_jose	0.001	213	0	174	0.004	179
austinmahone	0.002	143	0	126	0.188	73
azizansari	0.001	231	0	209	0.206	42
barackobama	0.011	18	0	169	0	204
bbcbusiness	0.003	78	0	241	0	242
berniesanders	0.002	134	0.01	47	0.149	133
bet	0.005	53	0	158	0	195
beyonce	0.005	57	0.013	36	0.203	47
bigsean	0.01	22	0	177	0	209
billboard	0.009	26	0.025	18	0.203	46
billclinton	0.002	124	0.003	97	0.184	75
billdeblasio	0.007	39	0.014	32	0.25	5
billgates	0.003	96	0.007	62	0.197	56
billpostmus	0.008	31	0.01	46	0.223	20
bishopnoeljones	0.003	98	0.005	71	0.235	12
blakeshelton	0.001	199	0	183	0	213

Continued on next page

Table 18 – continued from previous page

User	PageRank Score	PageRank Ranking	Betweenness Centrality	Betweenness Ranking	Closeness Centrality	Closeness Ranking
bobbyjindal	0.001	179	0	224	0	231
borisjohnson	0.003	79	0	207	0	221
britneyspears	0.001	232	0	210	0.012	166
brunomars	0.002	148	0	136	0.004	189
business	0.002	132	0.002	103	0.146	137
businessinsider	0.002	161	0	142	0.182	82
buzzfeed	0.002	154	0.013	37	0.193	61
bw	0.001	204	0	151	0.19	68
camerondallas	0.002	109	0.009	51	0.217	25
carlyfiorina	0.002	137	0.007	59	0.192	62
carmeloanthony	0.001	194	0	238	0.147	135
cbsnews	0.004	66	0.006	65	0.206	40
channingtatum	0.002	107	0	211	0	223
charliesheen	0.001	236	0	220	0.004	191
chrisbrown	0.003	89	0	225	0	232
chrischristie	0.002	122	0.006	64	0.178	87
ciara	0.001	229	0	205	0.135	152
cnbc	0.008	30	0	160	0	197
cnni	0.005	54	0	201	0	219
cnnmoney	0.002	123	0.004	79	0.134	154
cntraveler	0.002	140	0.027	17	0.212	35

Table 19: Retweets Rankings.

User	PageRank Score	PageRank Ranking	Betweenness Centrality	Betweenness Ranking	Closeness Centrality	Closeness Ranking
50cent	0.001	242	0.001	184	0.292	94
abc	0.004	64	0	246	0	267
actuallynph	0.001	232	0	229	0.231	195
alfranken	0.001	169	0.003	122	0.243	182
algore	0.001	197	0	208	0.286	104
aliciakeys	0.002	113	0.005	88	0.324	43
allenwest	0.001	236	0.003	120	0.243	184
ambassadorrice	0.001	260	0	200	0.263	145
andersoncooper	0.005	50	0	278	0	283
ap	0.001	243	0.002	158	0.301	77
aplusk	0.003	91	0.006	85	0.31	63
arianagrande	0.001	189	0.003	117	0.286	105
ashleytisdale	0.001	271	0	243	0.243	183
astro_jose	0.001	246	0	225	0.248	174
austinmahone	0.001	250	0	195	0.312	61
azizansari	0.008	29	0.022	20	0.312	62
barackobama	0.005	56	0.017	32	0.338	26
bbcbusiness	0.004	70	0.007	73	0.278	121
berniesanders	0.001	170	0.01	53	0.337	28
bet	0.02	5	0.035	9	0.306	68
beyonce	0.002	110	0.002	150	0.283	107
bigsean	0.002	114	0.013	46	0.302	76
billboard	0.001	258	0.004	103	0.289	98
billclinton	0.001	167	0.002	157	0.267	139

Continued on next page

Table 19 – continued from previous page

User	PageRank Score	PageRank Ranking	Betweenness Centrality	Betweenness Ranking	Closeness Centrality	Closeness Ranking
billdeblasio	0.001	208	0.001	183	0.228	200
billgates	0.001	267	0	233	0	259
billpostmus	0.001	162	0.005	96	0.326	39
bishopnoeljones	0.001	249	0	192	0.306	69
blakeshelton	0.007	41	0.003	131	0.003	249
bobbyjindal	0.002	126	0	280	0	285
borisjohnson	0.004	68	0.002	151	0.221	212
britneyspears	0.007	43	0	253	0.003	250
brunomars	0.001	283	0	273	0.234	190
business	0.001	190	0.003	116	0.264	140
businessinsider	0.001	187	0.001	163	0.261	149
buzzfeed	0.004	76	0	247	0	268
bw	0.001	175	0.005	95	0.322	46
camerondallas	0.005	54	0.002	161	0.28	111
carlyfiorina	0.001	164	0.003	140	0.32	48
carmeloanthony	0.001	264	0.004	101	0.278	118
cbsnews	0.001	216	0.003	115	0.333	32
channingtatum	0.006	45	0.008	70	0.184	230
charliesheen	0.009	28	0	196	0.003	245
chrisbrown	0.008	30	0	194	0.003	254
chrischristie	0.001	252	0	217	0.185	229
ciara	0.002	150	0.005	90	0.287	102
cnbc	0.002	152	0	264	0	278
cnni	0.001	287	0	284	0.225	206
cnnmoney	0.002	127	0.021	22	0.324	42
cntraveler	0.01	22	0.011	50	0.228	202

8.3.2. Mixing In the following tables, the number at index i, j corresponds to the average number of interactions in our networks originating from users in the category of row i and going towards the category of column j.

	Celebrities	**Media**	**Politicians**
Celebrities	128.9	38.72	10.64
Media	29.78	24.64	65.36
Politicians	5.63	60.15	214.53

Table 20: Average Number of Mentions Between User Categories.

In the following tables, the number at index i, j corresponds to the average percentage of interactions in our networks originating from users in the category of row i and directed towards the category of column j.

	Celebrities	Media	Politicians
Celebrities	10.11	1.47	0.48
Media	0.43	0.13	0.17
Politicians	0.21	0.68	4.13

Table 21: Average Number of Replies Between User Categories.

	Celebrities	Media	Politicians
Celebrities	19.40	14.75	2.05
Media	1.21	12.68	0.72
Politicians	0.75	15.09	35.7

Table 22: Average Number of Retweets Between User Categories.

8.4. Clustering

8.4.1. Cluster Characteristics In this section 26 we provide information about each cluster produced by each type affinity matrices we tested. The types of affinity matrices we tested are described in the "Choice of Affinity Matrix Construction Method" section 4.4.1.

Table 26: Cluster Info.

Cluster Method	Cluster Number	Conductance	Cluster Size	Num Celebrities	Num Media	Num Politicians
Mentions_2Clusters_Bib_Unweight	0	0.071	145	95	48	2
Mentions_2Clusters_Bib_Unweight	1	0.117	154	5	51	98

Continued on next page

	Celebrities	Media	Politicians
Celebrities	0.65	0.28	0.07
Media	0.29	0.32	0.39
Politicians	0.05	0.27	0.68

Table 23: Average Percentage of Mentions Between User Categories.

	Celebrities	Media	Politicians
Celebrities	0.79	0.15	0.06
Media	0.52	0.25	0.23
Politicians	0.08	0.14	0.78

Table 24: Average Percentage of Replies Between User Categories.

	Celebrities	Media	Politicians
Celebrities	0.45	0.48	0.06
Media	0.18	0.76	0.06
Politicians	0.05	0.36	0.59

Table 25: Average Percentage of Retweets Between User Categories.

Table 26 – continued from previous page

Cluster Method	Cluster Number	Conductance	Cluster Size	Num Celebrities	Num Media	Num Politicians
Mentions_2Clusters_Bib_Weight	0	0.22	195	23	73	99
Mentions_2Clusters_Bib_Weight	1	0.214	104	77	26	1
Mentions_2Clusters_Degdiscount_Unweight	0	0.071	145	95	48	2
Mentions_2Clusters_Degdiscount_Unweight	1	0.117	154	5	51	98
Mentions_2Clusters_Degdiscount_Weight	0	1.502	127	61	38	28
Mentions_2Clusters_Degdiscount_Weight	1	0.352	172	39	61	72
Mentions_2Clusters_Sum_Unweighted	0	0.122	147	3	46	98
Mentions_2Clusters_Sum_Unweighted	1	0.067	152	97	53	2
Mentions_2Clusters_Sum_Weighted	0	0.15	192	99	65	28
Mentions_2Clusters_Sum_Weighted	1	0.194	107	1	34	72
Mentions_2Clusters_Undir_Unweight	0	0.434	149	54	49	46
Mentions_2Clusters_Undir_Unweight	1	0.577	150	46	50	54
Mentions_2Clusters_Undir_Weight	0	0.532	139	40	47	52
Mentions_2Clusters_Undir_Weight	1	0.588	160	60	52	48
Mentions_3Clusters_Bib_Unweight	0	2.132	70	2	36	32
Mentions_3Clusters_Bib_Unweight	1	0.069	136	94	41	1
Mentions_3Clusters_Bib_Unweight	2	0.074	93	4	22	67
Mentions_3Clusters_Bib_Weight	0	0.233	113	18	48	47
Mentions_3Clusters_Bib_Weight	1	0.234	90	74	16	0
Mentions_3Clusters_Bib_Weight	2	1.766	96	8	35	53
Mentions_3Clusters_Degdiscount_Unweight	0	0.081	98	6	29	63
Mentions_3Clusters_Degdiscount_Unweight	1	0.068	130	92	37	1
Mentions_3Clusters_Degdiscount_Unweight	2	1.889	71	2	33	36
Mentions_3Clusters_Degdiscount_Weight	0	0.325	151	38	54	59
Mentions_3Clusters_Degdiscount_Weight	1	2.867	86	16	32	38
Mentions_3Clusters_Degdiscount_Weight	2	0.33	62	46	13	3
Mentions_3Clusters_Sum_Unweighted	0	0.206	58	0	11	47
Mentions_3Clusters_Sum_Unweighted	1	0.085	118	92	25	1
Mentions_3Clusters_Sum_Unweighted	2	0.539	123	8	63	52
Mentions_3Clusters_Sum_Weighted	0	0.188	106	1	34	71
Mentions_3Clusters_Sum_Weighted	1	0.254	183	92	62	29
Mentions_3Clusters_Sum_Weighted	2	0.226	10	7	3	0
Mentions_3Clusters_Undir_Unweight	0	0.638	106	39	32	35
Mentions_3Clusters_Undir_Unweight	1	0.604	108	36	40	32
Mentions_3Clusters_Undir_Unweight	2	0.75	85	25	27	33
Mentions_3Clusters_Undir_Weight	0	0.53	140	40	48	52
Mentions_3Clusters_Undir_Weight	1	0.576	150	57	49	44
Mentions_3Clusters_Undir_Weight	2	0.789	9	3	2	4
Mentions_4Clusters_Bib_Unweight	0	0.2	60	2	2	56
Mentions_4Clusters_Bib_Unweight	1	0.104	114	90	24	0
Mentions_4Clusters_Bib_Unweight	2	0.245	81	8	42	31
Mentions_4Clusters_Bib_Unweight	3	3.655	44	0	31	13
Mentions_4Clusters_Bib_Weight	0	2.066	91	8	32	51
Mentions_4Clusters_Bib_Weight	1	0.237	112	20	45	47
Mentions_4Clusters_Bib_Weight	2	0.293	81	64	15	2
Mentions_4Clusters_Bib_Weight	3	0.277	15	8	7	0
Mentions_4Clusters_Degdiscount_Unweight	0	0.301	93	82	11	0
Mentions_4Clusters_Degdiscount_Unweight	1	0.074	80	4	12	64
Mentions_4Clusters_Degdiscount_Unweight	2	2.183	65	2	30	33
Mentions_4Clusters_Degdiscount_Unweight	3	0.581	61	12	46	3
Mentions_4Clusters_Degdiscount_Weight	0	0.328	150	38	53	59

Continued on next page

Table 26 – continued from previous page

Cluster Method	Cluster Number	Conductance	Cluster Size	Num Celebrities	Num Media	Num Politicians
Mentions_4Clusters_Degdiscount_Weight	1	2.79	87	12	35	40
Mentions_4Clusters_Degdiscount_Weight	2	0.307	18	14	4	0
Mentions_4Clusters_Degdiscount_Weight	3	0.531	44	36	7	1
Mentions_4Clusters_Sum_Unweighted	0	0.342	40	31	8	1
Mentions_4Clusters_Sum_Unweighted	1	0.193	82	58	23	1
Mentions_4Clusters_Sum_Unweighted	2	0.448	121	11	58	52
Mentions_4Clusters_Sum_Unweighted	3	0.299	56	0	10	46
Mentions_4Clusters_Sum_Weighted	0	0.979	121	17	53	51
Mentions_4Clusters_Sum_Weighted	1	0.237	75	0	27	48
Mentions_4Clusters_Sum_Weighted	2	0.183	93	76	16	1
Mentions_4Clusters_Sum_Weighted	3	0.226	10	7	3	0
Mentions_4Clusters_Undir_Unweight	0	0.785	70	17	23	30
Mentions_4Clusters_Undir_Unweight	1	0.81	54	17	26	11
Mentions_4Clusters_Undir_Unweight	2	0.708	80	32	22	26
Mentions_4Clusters_Undir_Unweight	3	0.617	95	34	28	33
Mentions_4Clusters_Undir_Weight	0	0.81	109	38	38	33
Mentions_4Clusters_Undir_Weight	1	0.789	9	3	2	4
Mentions_4Clusters_Undir_Weight	2	0.556	123	43	40	40
Mentions_4Clusters_Undir_Weight	3	0.897	58	16	19	23
Replies_2Clusters_Bib_Unweight	0	0.081	124	92	30	2
Replies_2Clusters_Bib_Unweight	1	0.101	119	5	32	82
Replies_2Clusters_Bib_Weight	0	0.079	126	91	33	2
Replies_2Clusters_Bib_Weight	1	0.091	117	6	29	82
Replies_2Clusters_Degdiscount_Unweight	0	0.072	112	5	27	80
Replies_2Clusters_Degdiscount_Unweight	1	0.14	131	92	35	4
Replies_2Clusters_Degdiscount_Weight	0	0.092	130	92	34	4
Replies_2Clusters_Degdiscount_Weight	1	0.096	113	5	28	80
Replies_2Clusters_Sum_Unweighted	0	0.101	119	7	30	82
Replies_2Clusters_Sum_Unweighted	1	0.047	124	90	32	2
Replies_2Clusters_Sum_Weighted	0	0.091	118	6	29	83
Replies_2Clusters_Sum_Weighted	1	0.059	125	91	33	1
Replies_2Clusters_Undir_Unweight	0	0.47	121	39	37	45
Replies_2Clusters_Undir_Unweight	1	0.566	122	58	25	39
Replies_2Clusters_Undir_Weight	0	0.499	118	36	36	46
Replies_2Clusters_Undir_Weight	1	0.576	125	61	26	38
Replies_3Clusters_Bib_Unweight	0	0.085	127	91	33	3
Replies_3Clusters_Bib_Unweight	1	0.106	27	2	1	24
Replies_3Clusters_Bib_Unweight	2	0.263	89	4	28	57
Replies_3Clusters_Bib_Weight	0	1.216	85	3	27	55
Replies_3Clusters_Bib_Weight	1	0.079	123	91	31	1
Replies_3Clusters_Bib_Weight	2	0.149	35	3	4	28
Replies_3Clusters_Degdiscount_Unweight	0	0.097	138	93	42	3
Replies_3Clusters_Degdiscount_Unweight	1	3.6	52	1	15	36
Replies_3Clusters_Degdiscount_Unweight	2	0.084	53	3	5	45
Replies_3Clusters_Degdiscount_Weight	0	0.116	167	93	50	24
Replies_3Clusters_Degdiscount_Weight	1	0.152	39	3	4	32
Replies_3Clusters_Degdiscount_Weight	2	5.578	37	1	8	28
Replies_3Clusters_Sum_Unweighted	0	0.15	95	5	28	62
Replies_3Clusters_Sum_Unweighted	1	0.055	124	90	32	2
Replies_3Clusters_Sum_Unweighted	2	0.67	24	2	2	20
Replies_3Clusters_Sum_Weighted	0	0.061	123	91	31	1

Continued on next page

Table 26 – continued from previous page

Cluster Method	Cluster Number	Conductance	Cluster Size	Num Celebrities	Num Media	Num Politicians
Replies_3Clusters_Sum_Weighted	1	0.421	85	5	29	51
Replies_3Clusters_Sum_Weighted	2	0.264	35	1	2	32
Replies_3Clusters_Undir_Unweight	0	0.473	98	30	31	37
Replies_3Clusters_Undir_Unweight	1	0.557	120	58	25	37
Replies_3Clusters_Undir_Unweight	2	1.381	25	9	6	10
Replies_3Clusters_Undir_Weight	0	0.57	119	57	26	36
Replies_3Clusters_Undir_Weight	1	0.533	93	31	31	31
Replies_3Clusters_Undir_Weight	2	0.713	31	9	5	17
Replies_4Clusters_Bib_Unweight	0	0.087	125	90	32	3
Replies_4Clusters_Bib_Unweight	1	0.086	44	2	3	39
Replies_4Clusters_Bib_Unweight	2	4.547	50	2	26	22
Replies_4Clusters_Bib_Unweight	3	0.283	24	3	1	20
Replies_4Clusters_Bib_Weight	0	0.081	124	92	31	1
Replies_4Clusters_Bib_Weight	1	0.39	65	2	22	41
Replies_4Clusters_Bib_Weight	2	0.133	34	3	4	27
Replies_4Clusters_Bib_Weight	3	12.95	20	0	5	15
Replies_4Clusters_Degdiscount_Unweight	0	0.097	138	93	42	3
Replies_4Clusters_Degdiscount_Unweight	1	0.106	43	2	6	35
Replies_4Clusters_Degdiscount_Unweight	2	5.095	46	1	14	31
Replies_4Clusters_Degdiscount_Unweight	3	0.043	16	1	0	15
Replies_4Clusters_Degdiscount_Weight	0	0.116	167	93	50	24
Replies_4Clusters_Degdiscount_Weight	1	0.131	39	3	4	32
Replies_4Clusters_Degdiscount_Weight	2	7.378	22	0	5	17
Replies_4Clusters_Degdiscount_Weight	3	3.808	15	1	3	11
Replies_4Clusters_Sum_Unweighted	0	0.754	86	16	39	31
Replies_4Clusters_Sum_Unweighted	1	0.316	21	2	2	17
Replies_4Clusters_Sum_Unweighted	2	0.124	95	78	16	1
Replies_4Clusters_Sum_Unweighted	3	0.226	41	1	5	35
Replies_4Clusters_Sum_Weighted	0	0.305	31	1	1	29
Replies_4Clusters_Sum_Weighted	1	0.065	119	88	30	1
Replies_4Clusters_Sum_Weighted	2	1.472	18	2	2	14
Replies_4Clusters_Sum_Weighted	3	0.391	75	6	29	40
Replies_4Clusters_Undir_Unweight	0	0.576	82	24	29	29
Replies_4Clusters_Undir_Unweight	1	0.614	97	50	19	28
Replies_4Clusters_Undir_Unweight	2	1.303	22	9	6	7
Replies_4Clusters_Undir_Unweight	3	0.809	42	14	8	20
Replies_4Clusters_Undir_Weight	0	0.617	103	52	20	31
Replies_4Clusters_Undir_Weight	1	1.349	17	7	6	4
Replies_4Clusters_Undir_Weight	2	0.509	90	28	30	32
Replies_4Clusters_Undir_Weight	3	0.768	33	10	6	17
Retweets_2Clusters_Bib_Unweight	0	0.101	142	3	48	91
Retweets_2Clusters_Bib_Unweight	1	0.119	147	95	48	4
Retweets_2Clusters_Bib_Weight	0	0.118	160	10	55	95
Retweets_2Clusters_Bib_Weight	1	0.048	129	88	41	0
Retweets_2Clusters_Degdiscount_Unweight	0	0.117	142	4	47	91
Retweets_2Clusters_Degdiscount_Unweight	1	0.072	147	94	49	4
Retweets_2Clusters_Degdiscount_Weight	0	0.125	52	0	8	44
Retweets_2Clusters_Degdiscount_Weight	1	0.228	237	98	88	51
Retweets_2Clusters_Sum_Unweighted	0	0.083	168	97	63	8
Retweets_2Clusters_Sum_Unweighted	1	0.137	121	1	33	87
Retweets_2Clusters_Sum_Weighted	0	0.749	258	98	94	66

Continued on next page

Table 26 – continued from previous page

Cluster Method	Cluster Number	Conductance	Cluster Size	Num Celebrities	Num Media	Num Politicians
Retweets_2Clusters_Sum_Weighted	1	0.182	31	0	2	29
Retweets_2Clusters_Undir_Unweight	0	0.511	169	74	58	37
Retweets_2Clusters_Undir_Unweight	1	0.41	120	24	38	58
Retweets_2Clusters_Undir_Weight	0	0.859	259	92	88	79
Retweets_2Clusters_Undir_Weight	1	0.569	30	6	8	16
Retweets_3Clusters_Bib_Unweight	0	0.026	64	1	6	57
Retweets_3Clusters_Bib_Unweight	1	0.024	139	94	40	5
Retweets_3Clusters_Bib_Unweight	2	1.883	86	3	50	33
Retweets_3Clusters_Bib_Weight	0	0.183	113	7	39	67
Retweets_3Clusters_Bib_Weight	1	0.369	61	12	21	28
Retweets_3Clusters_Bib_Weight	2	0.143	115	79	36	0
Retweets_3Clusters_Degdiscount_Unweight	0	3.732	104	14	67	23
Retweets_3Clusters_Degdiscount_Unweight	1	0.077	80	2	9	69
Retweets_3Clusters_Degdiscount_Unweight	2	0.066	105	82	20	3
Retweets_3Clusters_Degdiscount_Weight	0	0.163	106	75	31	0
Retweets_3Clusters_Degdiscount_Weight	1	0.51	148	11	58	79
Retweets_3Clusters_Degdiscount_Weight	2	0.29	35	12	7	16
Retweets_3Clusters_Sum_Unweighted	0	0.1	60	0	11	49
Retweets_3Clusters_Sum_Unweighted	1	0.082	162	94	60	8
Retweets_3Clusters_Sum_Unweighted	2	0.206	67	4	25	38
Retweets_3Clusters_Sum_Weighted	0	0.408	168	29	71	68
Retweets_3Clusters_Sum_Weighted	1	0.116	27	0	0	27
Retweets_3Clusters_Sum_Weighted	2	0.234	94	69	25	0
Retweets_3Clusters_Undir_Unweight	0	0.87	85	17	25	43
Retweets_3Clusters_Undir_Unweight	1	0.421	106	42	32	32
Retweets_3Clusters_Undir_Unweight	2	0.743	98	39	39	20
Retweets_3Clusters_Undir_Weight	0	0.458	27	5	7	15
Retweets_3Clusters_Undir_Weight	1	0.836	87	33	34	20
Retweets_3Clusters_Undir_Weight	2	0.686	175	60	55	60
Retweets_4Clusters_Bib_Unweight	0	0.038	47	0	4	43
Retweets_4Clusters_Bib_Unweight	1	0.076	109	82	23	4
Retweets_4Clusters_Bib_Unweight	2	0.054	57	7	8	42
Retweets_4Clusters_Bib_Unweight	3	5.33	76	9	61	6
Retweets_4Clusters_Bib_Weight	0	0.122	108	77	30	1
Retweets_4Clusters_Bib_Weight	1	1.227	123	11	60	52
Retweets_4Clusters_Bib_Weight	2	0.193	30	10	6	14
Retweets_4Clusters_Bib_Weight	3	0.134	28	0	0	28
Retweets_4Clusters_Degdiscount_Unweight	0	0.141	83	65	15	3
Retweets_4Clusters_Degdiscount_Unweight	1	2.789	67	3	34	30
Retweets_4Clusters_Degdiscount_Unweight	2	0.097	77	4	12	61
Retweets_4Clusters_Degdiscount_Unweight	3	3.059	62	26	35	1
Retweets_4Clusters_Degdiscount_Weight	0	0.147	148	77	51	20
Retweets_4Clusters_Degdiscount_Weight	1	0.248	33	11	7	15
Retweets_4Clusters_Degdiscount_Weight	2	4.512	75	10	37	28
Retweets_4Clusters_Degdiscount_Weight	3	0.108	33	0	1	32
Retweets_4Clusters_Sum_Unweighted	0	0.108	90	69	21	0
Retweets_4Clusters_Sum_Unweighted	1	0.499	98	27	50	21
Retweets_4Clusters_Sum_Unweighted	2	0.144	48	0	8	40
Retweets_4Clusters_Sum_Unweighted	3	0.175	53	2	17	34
Retweets_4Clusters_Sum_Weighted	0	0.803	131	27	57	47
Retweets_4Clusters_Sum_Weighted	1	0.404	42	3	15	24

Continued on next page

Table 26 – continued from previous page

Cluster Method	Cluster Number	Conductance	Cluster Size	Num Celebrities	Num Media	Num Politicians
Retweets_4Clusters_Sum_Weighted	2	0.229	92	68	24	0
Retweets_4Clusters_Sum_Weighted	3	0.132	24	0	0	24
Retweets_4Clusters_Undir_Unweight	0	1.076	53	14	15	24
Retweets_4Clusters_Undir_Unweight	1	0.71	93	37	36	20
Retweets_4Clusters_Undir_Unweight	2	0.465	47	8	13	26
Retweets_4Clusters_Undir_Unweight	3	0.687	96	39	32	25
Retweets_4Clusters_Undir_Weight	0	0.526	129	47	40	42
Retweets_4Clusters_Undir_Weight	1	0.468	26	4	7	15
Retweets_4Clusters_Undir_Weight	2	1.254	47	13	16	18
Retweets_4Clusters_Undir_Weight	3	0.736	87	34	33	20

8.4.2. Cluster Evaluation Metrics In this section we provide evaluation metrics for clustering according to each of the affinity matrices we tested.

We define average conductance score, homogeneity, completenesss score, and v-measure score as described in the Evaluation subsection 5.4.2 of the Results section.

Table 27: Evaluation Metrics.

Cluster Method	Number of Clusters	Average Conductance	Homogeneity Score	Completeness Score	V-measure Score
Mentions_Bib_Unweight	2	0.095	0.331	0.526	0.407
Mentions_Bib_Weight	2	0.218	0.233	0.397	0.294
Mentions_Degdiscount_Unweight	2	0.095	0.331	0.526	0.407
Mentions_Degdiscount_Weight	2	0.84	0.036	0.058	0.044
Mentions_Sum_Unweighted	2	0.094	0.352	0.558	0.431
Mentions_Sum_Weighted	2	0.166	0.202	0.341	0.254
Mentions_Undir_Unweight	2	0.506	0.002	0.003	0.002
Mentions_Undir_Weight	2	0.562	0.005	0.007	0.006
Mentions_Bib_Unweight	3	0.554	0.357	0.37	0.363
Mentions_Bib_Weight	3	0.725	0.257	0.258	0.257
Mentions_Degdiscount_Unweight	3	0.505	0.33	0.34	0.335
Mentions_Degdiscount_Weight	3	1.057	0.099	0.105	0.102
Mentions_Sum_Unweighted	3	0.295	0.381	0.398	0.389
Mentions_Sum_Weighted	3	0.23	0.203	0.286	0.238
Mentions_Undir_Unweight	3	0.658	0.004	0.004	0.004
Mentions_Undir_Weight	3	0.56	0.006	0.008	0.007
Mentions_Bib_Unweight	4	0.684	0.456	0.378	0.413
Mentions_Bib_Weight	4	0.811	0.218	0.194	0.205
Mentions_Degdiscount_Unweight	4	0.707	0.461	0.37	0.41
Mentions_Degdiscount_Weight	4	1.073	0.14	0.133	0.136
Mentions_Sum_Unweighted	4	0.336	0.338	0.285	0.309
Mentions_Sum_Weighted	4	0.52	0.317	0.293	0.304
Mentions_Undir_Unweight	4	0.715	0.019	0.015	0.017
Mentions_Undir_Weight	4	0.722	0.004	0.004	0.004
Replies_Bib_Unweight	2	0.091	0.366	0.572	0.446

Continued on next page

Table 27 – continued from previous page

Cluster Method	Number of Clusters	Average Conductance	Homogeneity Score	Completeness Score	V-measure Score
Replies_Bib_Weight	2	0.085	0.355	0.555	0.433
Replies_Degdiscount_Unweight	2	0.109	0.34	0.533	0.415
Replies_Degdiscount_Weight	2	0.094	0.34	0.533	0.415
Replies_Sum_Unweighted	2	0.073	0.345	0.54	0.421
Replies_Sum_Weighted	2	0.074	0.371	0.58	0.452
Replies_Undir_Unweight	2	0.518	0.012	0.019	0.015
Replies_Undir_Weight	2	0.539	0.017	0.026	0.02
Replies_Bib_Unweight	3	0.152	0.363	0.413	0.386
Replies_Bib_Weight	3	0.486	0.383	0.419	0.4
Replies_Degdiscount_Unweight	3	0.844	0.385	0.424	0.403
Replies_Degdiscount_Weight	3	0.954	0.209	0.27	0.235
Replies_Sum_Unweighted	3	0.153	0.356	0.41	0.381
Replies_Sum_Weighted	3	0.216	0.397	0.434	0.415
Replies_Undir_Unweight	3	0.608	0.015	0.017	0.016
Replies_Undir_Weight	3	0.574	0.022	0.024	0.023
Replies_Bib_Unweight	4	1.024	0.4	0.359	0.378
Replies_Bib_Weight	4	1.23	0.398	0.366	0.382
Replies_Degdiscount_Unweight	4	1.041	0.393	0.379	0.385
Replies_Degdiscount_Weight	4	1.004	0.212	0.244	0.227
Replies_Sum_Unweighted	4	0.381	0.351	0.305	0.327
Replies_Sum_Weighted	4	0.301	0.385	0.357	0.371
Replies_Undir_Unweight	4	0.697	0.027	0.023	0.025
Replies_Undir_Weight	4	0.649	0.027	0.024	0.025
Retweets_Bib_Unweight	2	0.11	0.327	0.518	0.401
Retweets_Bib_Weight	2	0.087	0.318	0.508	0.391
Retweets_Degdiscount_Unweight	2	0.094	0.316	0.502	0.388
Retweets_Degdiscount_Weight	2	0.209	0.136	0.316	0.19
Retweets_Sum_Unweighted	2	0.106	0.32	0.517	0.396
Retweets_Sum_Weighted	2	0.689	0.095	0.308	0.146
Retweets_Undir_Unweight	2	0.469	0.043	0.07	0.053
Retweets_Undir_Weight	2	0.829	0.01	0.033	0.015
Retweets_Bib_Unweight	3	0.578	0.38	0.398	0.389
Retweets_Bib_Weight	3	0.206	0.272	0.281	0.276
Retweets_Degdiscount_Unweight	3	1.388	0.39	0.393	0.392
Retweets_Degdiscount_Weight	3	0.356	0.265	0.302	0.282
Retweets_Sum_Unweighted	3	0.114	0.304	0.338	0.32
Retweets_Sum_Weighted	3	0.324	0.282	0.343	0.309
Retweets_Undir_Unweight	3	0.662	0.035	0.035	0.035
Retweets_Undir_Weight	3	0.71	0.017	0.021	0.019
Retweets_Bib_Unweight	4	1.447	0.442	0.364	0.399
Retweets_Bib_Weight	4	0.601	0.323	0.298	0.31
Retweets_Degdiscount_Unweight	4	1.369	0.36	0.287	0.319
Retweets_Degdiscount_Weight	4	1.287	0.188	0.174	0.181
Retweets_Sum_Unweighted	4	0.259	0.331	0.272	0.299
Retweets_Sum_Weighted	4	0.507	0.282	0.256	0.269
Retweets_Undir_Unweight	4	0.73	0.037	0.03	0.033
Retweets_Undir_Weight	4	0.703	0.021	0.018	0.019

8.4.3. Cluster Assignments In this section we provide the cluster assignments produced by forming 3 clusters from the network of mentions. Cluster assignments for other clustering methods are available at `https://github.com/cchen23/Fall-2016-IW/tree/master/data/clusters`.

Table 28: Cluster Assignment, 3 Clusters, Mentions.

Name	Cluster	Category	Name	Cluster	Category	Name	Cluster	Category
abc	0	m	alfranken	0	p	algore	0	p
allenwest	0	p	ambassadorrice	0	p	andersoncooper	0	c
ap	0	m	astro_jose	0	p	barackobama	0	p
bbcbusiness	0	m	berniesanders	0	p	billclinton	0	p
billdeblasio	0	p	billpostmus	0	p	bishopnoeljones	0	p
bobbyjindal	0	p	borisjohnson	0	p	business	0	m
businessinsider	0	m	bw	0	m	carlyfiorina	0	p
cbsnews	0	m	chrischristie	0	p	cnbc	0	m
cnni	0	m	cnnmoney	0	m	condoleezzarice	0	p
corybooker	0	p	darrellissa	0	p	davidaxelrod	0	p
dccc	0	p	donnabrazile	0	p	dwstweets	0	p
elizabethforma	0	p	ericcantor	0	p	ericgarcetti	0	p
foramerica	0	p	fortunemagazine	0	m	foxnews	0	m
freedomworks	0	p	gabbygiffords	0	p	gameinformer	0	m
gavinnewsom	0	p	georgehwbush	0	p	gop	0	p
gov	0	m	govchristie	0	p	governorperry	0	p
govgaryjohnson	0	p	govmikehuckabee	0	p	govwalker	0	p
harvardbiz	0	m	hillaryclinton	0	p	history	0	m
housedemocrats	0	p	inc	0	m	jasoninthehouse	0	p
jebbush	0	p	jeffarazzi	0	p	jerrybrowngov	0	p
jimcarrey	0	c	jimdemint	0	p	joebiden	0	p
joetrippi	0	p	johnemichel	0	p	johnkasich	0	p
johnkerry	0	p	johnkingated	0	p	karlrove	0	p
kasimreed	0	p	lagarde	0	p	latimes	0	m
life	0	m	lindasuhler	0	p	madam_president	0	p
marcorubio	0	p	marketwatch	0	m	massgovernor	0	p
mcfaul	0	p	michelebachmann	0	p	mittromney	0	p
mufihannemann	0	p	nancypelosi	0	p	nasahqphoto	0	p
natgeophotos	0	m	natgeotravel	0	m	nbcnews	0	m
newsweek	0	m	newtgingrich	0	p	newyorker	0	m
nikkihaley	0	p	npr	0	m	nprpolitics	0	m
nytimes	0	m	nytimesbooks	0	m	nytimesworld	0	m
pbs	0	m	politico	0	m	potus	0	p
preetbharara	0	p	pryan	0	p	randpaul	0	p
rbreich	0	p	realbencarson	0	p	realdonaldtrump	0	p
reince	0	p	repjohnlewis	0	p	repronpaul	0	p
ricksantorum	0	p	ronpaul	0	p	sarahpalinusa	0	p
sciam	0	m	scottwalker	0	p	senatorreid	0	p
sengillibrand	0	p	senjohnmccain	0	p	senmikelee	0	p
sensanders	0	p	senschumer	0	p	sentedcruz	0	p
senwarren	0	p	slate	0	m	speakerboehner	0	p
speakerryan	0	p	stephenathome	0	c	teamcantor	0	p
tedcruz	0	p	tgowdysc	0	p	theatlantic	0	m

Continued on next page

Table 28 – continued from previous page

Name	Cluster	Category		Name	Cluster	Category		Name	Cluster	Category
thedemocrats	0	p		thehermancain	0	p		thehill	0	m
thenextweb	0	m		thereval	0	p		timkaine	0	p
tppatriots	0	p		twitterapi	0	m		uninoticias	0	m
univision	0	m		usatoday	0	m		voterocky2016	0	p
washingtonpost	0	m		weatherchannel	0	m		wendydavis	0	p
wired	0	m		wiredscience	0	m		wsj	0	m
50cent	1	c		actuallynph	1	c		aliciakeys	1	c
aplusk	1	c		arianagrande	1	c		ashleytisdale	1	c
austinmahone	1	c		azizansari	1	c		bet	1	m
beyonce	1	c		bigsean	1	c		billboard	1	m
billgates	1	c		blakeshelton	1	c		britneyspears	1	c
brunomars	1	c		buzzfeed	1	m		camerondallas	1	c
carmeloanthony	1	c		channingtatum	1	c		charliesheen	1	c
chrisbrown	1	c		ciara	1	c		cntraveler	1	m
comedycentral	1	m		conanobrien	1	c		cosmopolitan	1	m
danieltosh	1	c		ddlovato	1	c		digg	1	m
discovery	1	m		ellemagazine	1	m		eminem	1	c
enews	1	m		entrepreneur	1	m		espndeportes	1	m
espnstatsinfo	1	m		esquirenetwork	1	m		evalongoria	1	c
ew	1	m		fastcompany	1	m		foodandwine	1	m
foxsports	1	m		gadgetlab	1	m		goodhealth	1	m
hbo	1	m		iamdiddy	1	c		iamqueenlatifah	1	c
iamwill	1	c		iggyazalea	1	c		jcolenc	1	c
jessicaalba	1	c		jessicasimpson	1	c		jimmyfallon	1	c
jimmykimmel	1	c		jlo	1	c		joejonas	1	c
johncena	1	c		johnlegend	1	c		jtimberlake	1	c
justinbieber	1	c		kanyewest	1	c		katyperry	1	c
kdtrey5	1	c		kelly_clarkson	1	c		kendalljenner	1	c
kendricklamar	1	c		kevinhart4real	1	c		khloekardashian	1	c
kimkardashian	1	c		kingjames	1	c		kobebryant	1	c
kourtneykardash	1	c		krisjenner	1	c		kyliejenner	1	c
ladygaga	1	c		leodicaprio	1	c		lifehacker	1	m
liltunechi	1	c		lindsaylohan	1	c		lmfao	1	c
louisfarrakhan	1	p		ludacris	1	c		lukebryanonline	1	c
marcanthony	1	c		mariahcarey	1	c		marieclaire	1	m
maroon5	1	c		marvel	1	m		mashable	1	m
menshealthmag	1	m		mileycyrus	1	c		mindykaling	1	c
mirandacosgrove	1	c		nbaontnt	1	m		nbatv	1	m
nbc	1	m		nflnetwork	1	m		nickelodeontv	1	m
nickiminaj	1	c		nickjonas	1	c		nytimesarts	1	m
oprah	1	c		parenting	1	m		parentsmagazine	1	m
parishilton	1	c		people	1	m		pharrell	1	c
pink	1	c		pitbull	1	c		realsimple	1	m
robertdowneyjr	1	c		rollingstone	1	m		ryanseacrest	1	c
sai	1	m		salahitareq	1	p		sarahksilverman	1	c
sciencechannel	1	m		selenagomez	1	c		serenawilliams	1	c
sethmacfarlane	1	c		shaq	1	c		shawnmendes	1	c
snoopdogg	1	c		stephencurry30	1	c		stevemartintogo	1	c
taylorswift13	1	c		techcrunch	1	m		teenvogue	1	m
theellenshow	1	c		therock	1	c		thr	1	m
tmagazine	1	m		tmz	1	m		tomhanks	1	c

Continued on next page

Table 28 – continued from previous page

Name	Cluster	Category		Name	Cluster	Category		Name	Cluster	Category
travelchannel	1	m		travelleisure	1	m		treysongz	1	c
tyga	1	c		tyrabanks	1	c		usher	1	c
usweekly	1	m		vanityfair	1	m		victoriajustice	1	c
voguemagazine	1	m		whowhatwear	1	m		wizkhalifa	1	c
womenshealthmag	1	m		wwd	1	m		xtina	1	c
younghollywood	1	m		zacefron	1	c				