CS224W Final Report
Existence of Pseudo-Local Information Diffusion Catalysts on Twitter

Dimuth Kulasinghe, Ashwin Apte (Group 18)

December 10, 2012

1 Introduction

1.1 Motivation

Our project aims to investigate mechanisms of information diffusion across the Twitter network in order to enable detection of rumor origination and propagation. Twitter provides an ideal platform for efficient dispersal of information across a vast audience in real time. A possible misuse of this medium, however, is the spread of malicious or untruthful information; as was recently reported to have incited violence and riots in India [4].

The size of the Twitter network makes it impossible for a hypothetical peacekeeper to keep surveillance of all individual nodes to detect rumors and prevent dissemination. We hypothesize, that the Twitter network can be mined to reveal underlying structures containing specific properties, where the structures can be analyzed to further reveal specific nodes which act as the primary catalysts of information diffusion across the network. This process could enable security agencies to discover and monitor the select nodes in real time.

1.2 Problem Definition

Our problem is to deduce whether there exist key nodes within underlying structures in a Twitter network that act as catalysts for information diffusion. We define underlying structures in the Twitter network as communities, and will partition the network into communities based on existing community-detection algorithms. Our basis for detecting information diffusion will be in the form of retweets: if user A retweets a tweet from user B, then we claim that B influenced A. A key node in the network, then, will be one that influences a high number of nodes. If we can find structural properties in a set of communities that allow us to filter out their most influential nodes, then we will have a process to systematically detect information catalysts in a large network.

1.3 Prior Work

We have not come across any other prior work which analyses a large social network (web or real world) to find communities and then studies rumor propagation or influencers in the communities. We describe here in brief related works that we have used as reference for our project.

Qazvinian et al. [5] attempt to provide a mechanism for automatic identification of rumors in social media through the use of Rumor Retrieval (extract tweets that highlight controversial aspects

of a given story, and spread misinformation), and Belief Classification (identify which followers believe in the misinformation). The model had a mean average precision of .95, and the authors concluded that rumors that have been identified as such by other means can be much more readily detected in Twitter feeds.

A paper by Cha and Haddadi [reference] explores the influence of a Twitter user based on three key indicators, viz. in- degree (number of followers), re-tweets and mentions. The paper argues quite convincingly that high in-degree indicates popularity of an individual, but by itself does not indicate engagement with the audience. Ranking by re-tweets is an indicator of strong content, and can be construed as a good proxy for how convincing and believable the followers and the tweeter. Ranking by number of mentions indicates the brand value of a user more than strength of the content.

de Quincey and Kostkova [3] describe the use Twitter to detect disease outbreaks, with the particular example of swine flu. They recommend using their results as extensions to existing epidemic intelligence tools.

A paper by Kewlaramani [2] attempts to find communities based on a variety of similarity metrics based on content of tweets, links, metadata and user profiles. The paper concludes that link similarity is a better indicator of community structure than content similarity.

2 Method

2.1 Data Collection

We obtained inter-related data from three sources, the first being a 67 GB collection of tweets harvested from 2009-06 to 2009-12 [*http://snap.stanford.edu/data/bigdata/*], the second being a 30 GB follower edgelist representing the same timespan [private source], and a 20 MB file describing the mappings between usernames and node ids [private source]. We scaled down the follower edgelist to only include nodes existing in the third source, and performed lexical analysis to extract retweet information from the first source. Retweet data was also filtered down to only include nodes from the third source. Finally, we used the Twitter API to retrieve location and language information on the users from the third source.

We used the language information to perform an initial geographic partitioning of the social network. We could not accurately perform a further partitioning using the location information, which consisted of a user entered description of location; the location descriptions were inconsistent and often meaningless. The English speaking community network was still too large to run community-detection algorithms on; the partitions ended up top-heavy, with a single community containing over 99% of the nodes in the subcommunity (as is expected in densely overlapping communities). We chose to do our primary analysis on the three largest non-English speaking partitions in our data, being Spanish, Japanese, and Portuguese, each consisting of about 7000 users. We refer to the three communities by their native country names from here on for convenience.

2.1.1 Community Detection

We use a paper by Lancichinetti and Fortunato [1] as the basis for algorithm selection. The paper compares 12 community detection algorithms using 2 benchmarks, namely the Girvan Newman (GN) Benchmark and the LFR Benchmark. The authors conclude that the InfoMAP method by Rosvall and Bergstrom performs best on the set of benchmarks tested in the paper. InfoMAP shows excellent results on the GN benchmark, as well as the much more rigorous LFR benchmark. Also, InfoMAP shows good performance on weighted and directed graphs as well. InfoMAP was tested for different sizes of network and communities, with uniform results. The running time for InfoMAP is linear in

network size, and hence can be used with graphs containing millions of nodes and links. We used the InfoMAP algorithm to partition each of these language partitions into further subcommunities.

The algorithm generalizes a flow based clustering method called the map equation to detect multiple levels in a network, and the relationships between the sub-modules at each level. The generalized map equation described in the paper attempts to answer 3 questions: a) How many hierarchical levels can the network be partitioned into? b) How many sub-modules at each hierarchical level? C) Which node belongs to which sub-module? The algorithm as described in the paper divides the network into hard partitions, ignoring possibility of overlap of nodes across partitions.

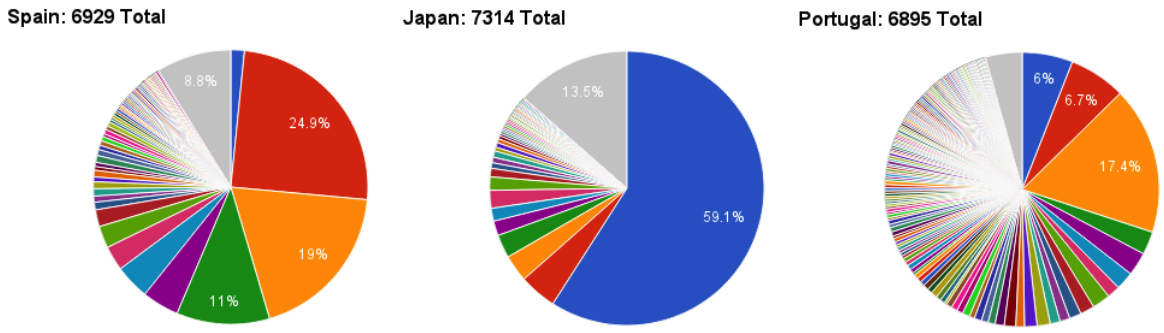

Figure 1: InfoMAP community partition distributions for each country.

Fig. 1 shows that all three partitions resulted in a major community which contained a comparatively high percentage of the country's members; this is especially true in the case of Japan. Nonetheless, these partitions are much more uniform than the extreme top heavy partitions resulting from running community detection on the entire network.

2.2 Analysis Guidelines

We use three traditional node attributes as our independent variables to measure node influence: Indegree, Pagerank, and Betweenness Centrality (referred to as Betweenness from here on). Attributes are calculated in respect to the nodes' community, not in respect to the country partition. We investigate whether there exists a correlation between a given attribute and a node's influence, and compare and contraste the correlations.

Pagerank can be viewed to be a measure of peer recognition of a node's influence. High Pagerank indicates nodes trusted by a large proportion of other trusted nodes.

Nodes with high Betweenness may be identified as *information brokers*. They have a relatively high outdegree, indicating access to a large amount of information, and a high indegree, which means a large audience for information dispersal. They lie on a large number of shortest paths, which may be viewed as a potential short circuit on the information pipeline, allowing for efficent information dispersal. These nodes thus may be in a position to play a powerful role in inciting as well as deterring rumor propagation and violence.

We hypothesize that Pagerank and Betweeness would prove to be the strongest indicators of influence in communities.

Since the specific values of these attributes are highly dependent on the graphs to which they are being applied, we focus not on the attribute values of the nodes, but on the *relative orderings* enforced by the attributes. We hypothesize that sorting a set of nodes by descending order of these

attributes will order the most influential nodes to the top; the actual attribute values are not so important.

We measure node influence on both a local and global scale; this is to see how often locally influential nodes are influential in respect to the entire network. The local influence of a node is measured as the percentage of nodes influenced in its community. As a node is unlikely to influence a significant percentage of nodes on a global scale, we measure the global influence of a node as a direct tally of the times it is retweeted. Global tallies take into account the entire dataset, not just the respective country. Tallies of retweets within communities are also calculated, but we choose not to report them in the paper as they mirror the influence metric described here.

We perform our analysis and display data on the level of countries. Each analysis metric is applied to the specific communities in the country partitions, and the resultant statistics are aggregated as a weighted sum based on community size.

Finally, a major challenge in our investigation is the absence of ground truth to compare our results against. At each stage, we have been forced to move ahead with only an intuitive guideline of what seems to be working well. Hence, our comparisons are relative and do not reference a concrete value.

2.3 Process

We use several metrics to determine correlations between the attribute values and node influence.

The first metric is an analysis of each attribute's *influence distribution*, which is an ordering of the influence values of the nodes for each country based on the attribute used; this will give a general visualization of the effectiveness of each attribute ordering. We will use as a benchmark an *ideal influence distribution*, where the node sets are intentionally ranked in order of influence. We show below the ideal distribution graph for Spain. The other 2 partitions also show similar graphs.

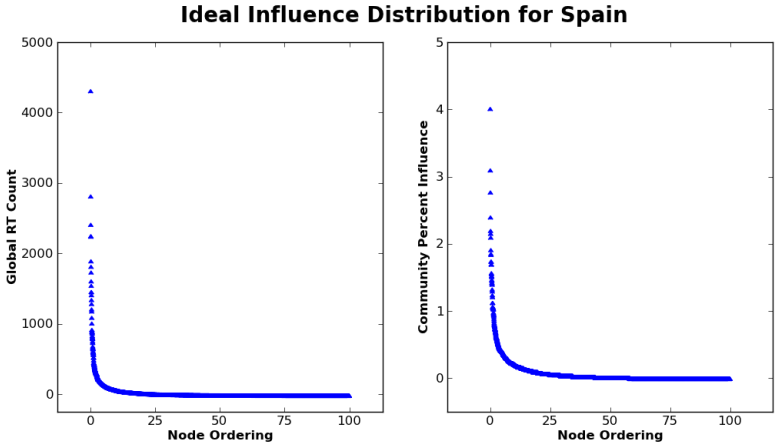

Figure 2: Ideal Distribution. y axis measures influence, x axis the normalized rank of the node within the country.

The ideal distributions follow an exponential trend; we want to measure the conformance of each attribute's influence distribution to its corresponding ideal distribution. For the log values of each ideal distribution, we fit a polynomial regression in the form

$$\log y = \log A_0 + A_1 n + A_2 n^2 + ... + A_{10} n^{10},$$

and calculated the sum of square error $SS_{error-ideal}$ for each. We then calculate the sum of square error $SS_{error-actual}$ for each of the actual influence distributions in comparison to the regression for its

ideal distribution, and return the ratio of its error over the ideal error ratio, $SS_{error_actual}/SS_{error-ideal}$. Finally, we compare these ratios to see which are the smallest.

The second metric is an analysis of the cumulative influence distributions, where we plot the cumulative percentage of total retweets attributed to the current node and all nodes of higher rank. We see in Fig. 2 that the most influential nodes in each graph occur well within top 25% ranked nodes. We compare the growth rates of each cumulative distribution over top 25% of its nodes by fitting a polynomial regression of degree three over the first quartile, and then comparing the derivatives of each of these polynomials. Finally, we take the percentage of influence values accumulated over the top 25% of each distribution, and plot as a ratio this value over percentage accumulated by the distribution's corresponding cumulative ideal distribution; this gives us a measure of ratio of the desired influence values each attribute manages to extract.

The third metric investigates the community inclusion statistics for each attribute. Since we are viewing our correlations from a country level, we want to see the participation percentage of communities in the higher ranked nodes. Thus, we take the top 25% ranked nodes in each distribution from metric 1, and measure the percentage of communities having contributed at least one node in the top x% of nodes. Also, for the top 10% of nodes from each metric 1 graph, we plot the distribution of the nodes as a measure of their respective community size. Ideally, we would like the top x% of nodes to come as a weighted proportion from the communities, where community with size s contributes sx/N nodes, where N is the number of nodes in the country.

3 Results

3.1 Metric 1: Influence Distribution

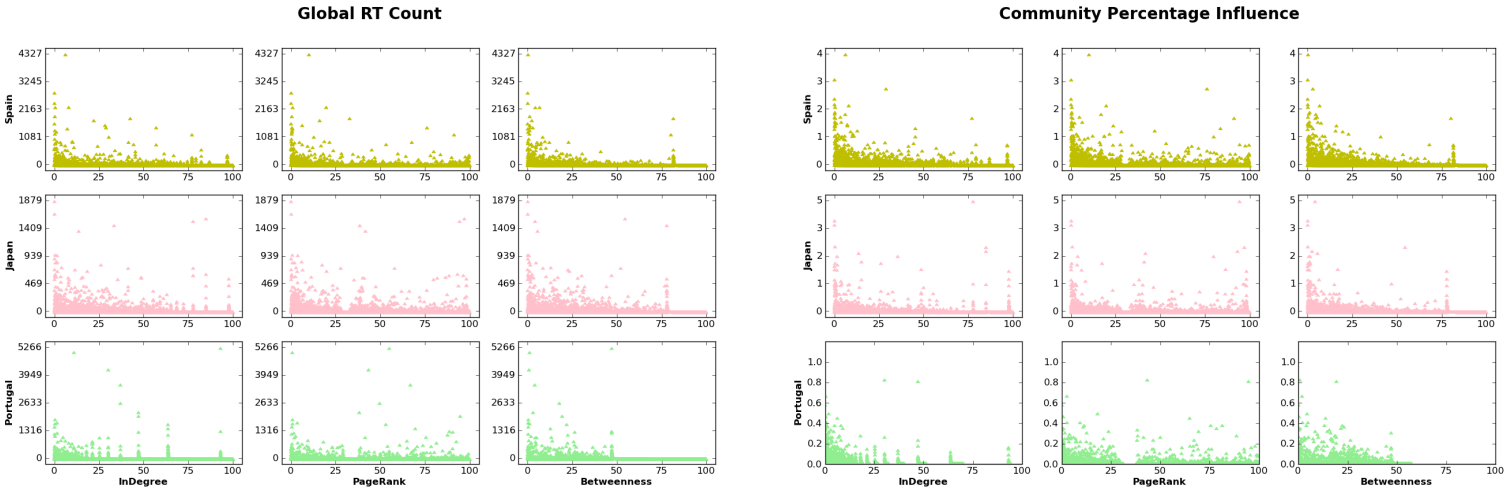

Figure 3: Global RT Count & Community Percentage Influence for each attribute

All three attribute distributions are visually comparable to the ideal distribution for each of the 3 countries, and depict prominent correlations with the global influence measures. Notably, the most highly influential nodes tend to be the top ranking nodes for each attribute. Another notable observation is that there are some scattered high influence nodes in the 3rd and 4th quartile for Indegree and Pagerank. Betweeness distributions, on the other hand, seem to effectively cut off high influence values past a certain threshold.

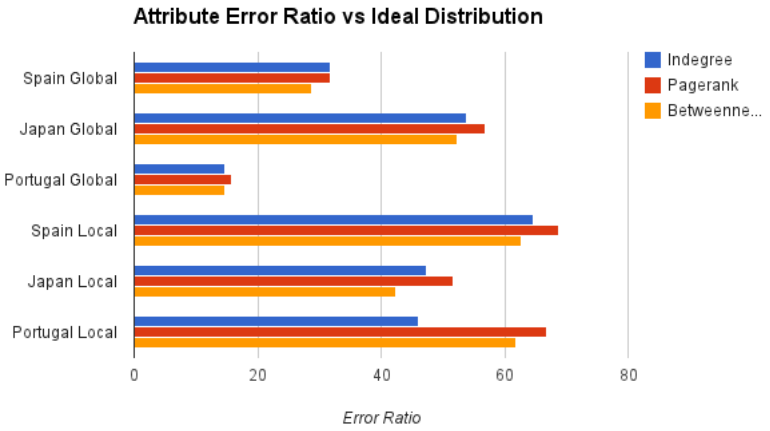

Figure 4: Attribute distribution error factor over ideal distribution factor (Conformance)

Pagerank is the highest ranking attribute in all the graphs. Each distribution regardless shows a high ratio, with factors well above 10; this implies that the distributions themselves do not match the trendlines of their ideal distributions very accurately.

Another observation from the Influence Distribution and Conformance graphs shows that all 3 attributes impose a much better ordering on the local community influence graphs for Spain and Portugal. Japan proves to be the exeption to this, where all 3 attributes perform better on ordering nodes for global retweet counts.

3.2 Metric 2: Cumulative Influence Distribution

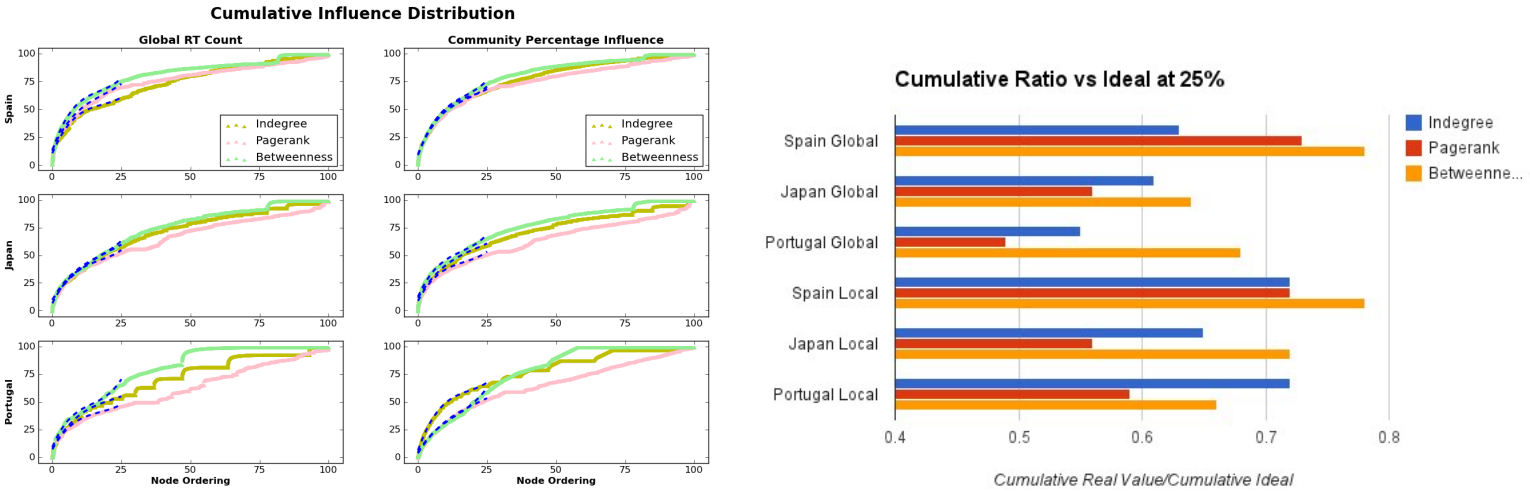

Figure 5: Cumulative Influence Distribution and Cumulative Observed vs Ideal Ratio

Cumulative influence distributions show a steep upward sloping curve, flattening out after the first 10% ordered nodes. The graphs show that between 50 and 75% of the global retweet count and influence metric comes from the top 25% of nodes. This roughly approaches Pareto's law, where the top 20% of nodes ranked by our metrics account for almost 80% of retweets. The Cumulative Observed vs. Ideal ratio is high for Betweenness across almost all categories of country - influence

graph pairs. Indegree and Pagerank show mixed results on this metric. Overall, the ratio is above 0.5 for all attributes, and above 0.7 in some cases. A high ratio indicates that the ordering approaches the ideal ordering (Fig. 2).

3.3 Metric 3: Community Inclusion

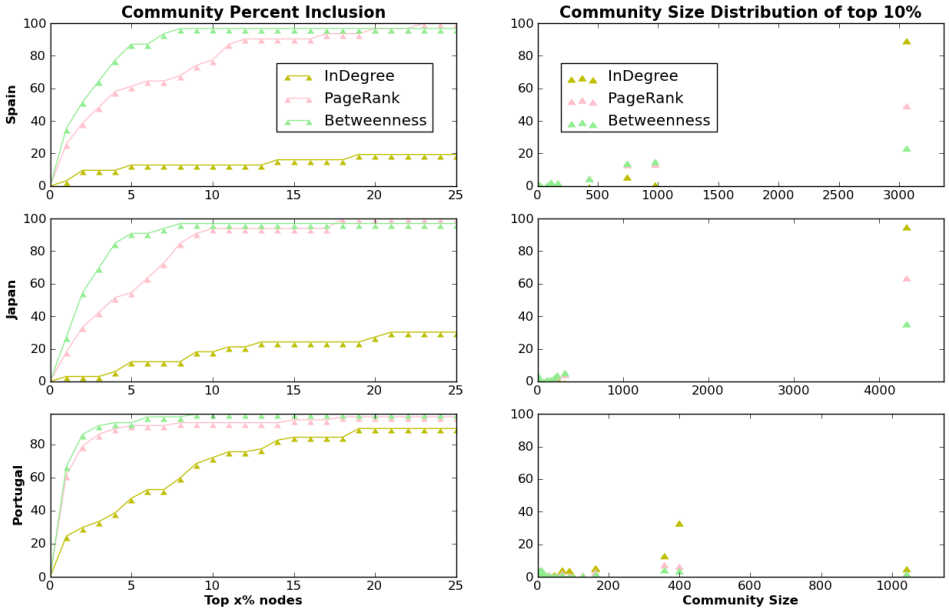

Figure 6: Cumulative Influence Distribution and Cumulative Observed vs Actual Ratio

Finally, we analyze how many communities and of what size contribute to the top ranked nodes for each attribute. We see that for Pagerank and Betweeness, almost all communities contribute to the top 10% nodes. Indegree on the other hand shows variability across countries. This behavior is explained by the community size distribution graph for top 10% nodes. A small number of large communities contribute for Indegree, indicating a size bias. Larger communites may have nodes with high indegree precisely because the community is larger. Pagerank and Betweenness show a better distribution of community size of included communities.

4 Conclusion

Our analysis throws up several interesting points with regard to identifying influential nodes within communities. We see that Indegree, Pagerank and Betweeness are all able to identify high influence nodes in communities. While conformance ratios to ideal distributions were undesirable, each attribute accumulated a high percentage of the total possible influence values within the first quartile of node orderings. These attributes of a node may thus be useful to identify potential threats for rumor origination and propagation.

Indegree suffers from a community size bias, due to which it may suppress influential nodes from smaller communities due to presence of larger communities. This may be because indegree is a short-sighted statistic that does not look past a node's immediate neighbors. This attribute may therefore be treated with caution when identifying influential nodes.

Pagerank seems to perform relatively well on Influence Distribution and Community Inclusion Metrics. It did not perform as well as the Betweenness attribute, however; we hypothesize this may be due to the definition of Pagerank in the Twitter context. The follower network may be an insufficient parameter through which to calculate Pagerank in this context, and may even be more useful taking into account the notion of retweet counts, making the analysis on this section circular.

Betweenness perfoms exceptionally well on Cumulative Influence Distribution and reasonably well on Community Inclusion and Influence Distribution metrics. Betweenness is entirely dependent on the node's incoming and outgoing edges, and this might be the reason it outperforms InDegree and Pagerank in our analysis. The performance of Betweenness indicates that this may be the most reliable attribute to discover influential nodes in a community.

We conclude that all three attributes, especially Betweenness, can be used as indicators of influence in Twitter networks. These attributes may be very powerful tools in the hands of security agencies for identifying potential sources of rumors, as well as effectively dealing with explosive situations.

4.1 Future Work

Our work should be viewed with the caveat that we worked with limitations on the size of data we could analyze, due to limitations on availabe computing resources. A similar study on a much larger sample of the network, or even the entire network would give more credence to our results.

Partitioning the network on other parameters such as location, religion, political ideology etc. would also be a direction to explore, as we see strong real world communities along these paramters. Our results would ideally hold across these partitons as well, providing us with the ideal tools to discover influence in communities.

Another direction for further research might be studying other measures of influence apart from retweet counts, such as mentions and indegree.

References

[1] Lancichinetti, Andrea, and Santo Fortunato. "Community Detection Algorithms: A Comparative Analysis." (n.d.): n. pag. *http://arxiv.org.* 16 Sept. 2010. Web. 15 Nov. 2012.

[2] Kewalramani, Mohit. "Community Detection in Twitter." Diss. University of Maryland Baltimore County, 2011. http://ebiquity.umbc.edu, 5 May 2011. Web. 15 Nov. 2012.

[3] Quincy, Ed, and Patty Kostova. "Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter." *http://users.cs.fiu.edu.* City EHealth Research Centre, n.d. Web.

[4] "Facebook, Twitter Accounts Blocked over Hate Messages." The Times of India. N.p., 20 Aug. 2012. Web. <http://articles.timesofindia.indiatimes.com/2012-08-20/internet/33287010_1_objectionable-contents-twitter-accounts-rumours>.

[5] Cha, M., & Haddadi, H. (2010). Measuring user influence in Twitter: The million follower fallacy. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)* (May 2010)