International Journal of Physical Sciences Vol. 7(13), pp. 2038 - 2051, 23 March, 2012
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.1280
ISSN 1992 - 1950 ©2012 Academic Journals

*Full Length Research Paper*

# Mapping the structure, evolution and geo-spatiality of a social media network

**Sameer Kumar[1]* and Jariah Mohd. Jan[2]**

[1]Asia-Europe Institute, University of Malaya, 50603 Kuala Lumpur, Malaysia.
[2]Faculty of Languages and Linguistics, University of Malaya, 50603 Kuala Lumpur, Malaysia.

Accepted 27 February, 2012

**Social media is the baby born out of the confluence of digital technology and human beings' desire to collaborate. Past researches in social media networks have mostly concentrated on investigation of large networks, which do not fully capture the micro-level dynamics of the network. In this study, an in-depth topological analysis of a small network (n=200) formed on Twitter during a 24 h period was carried out. The results showed that the network had both small-world and scale-free characteristics. Geo-spatiality revealed more interest by users in regions where the subject of tweets had its stake. The most influential nodes were those whose tweets got re-tweeted the most. Temporal analysis showed faster formation of network when there was a tweet of interest. Traditional news media had a powerful hold on the tweets being made by users. Communities formed around tweets of a certain theme and there was a common theme that kept the entire network together.**

**Key words:** Twitter, social media analysis, NodeXL, social network analysis.

## INTRODUCTION

Social Media has opened a new chapter in human beings' freedom of speech and action. People now freely collaborate, share videos, photos, news, reviews, opinions and stories using this media. Social networking sites like Facebook and Twitter facilitate individuals to connect with friends and acquaintances and remain in touch with them for as long as they wish. Researchers have been studying how broadcast of information on Twitter affect our thinking process, business and society as a whole (Chen, 2011; Chew and Eysenbach, 2010; Johnson, 2011; Ye and Wu, 2010). In 140 words or less, Twitter makes it possible for individuals to send short messages to those who follow them. This micro-blogging application, which now has over 200 million users (March 2011 estimates), is being increasingly used as an alternative to face-to-face interaction.

Profile and linkage data from Twitter can be collected using automated collection techniques enabling network researchers to understand the patterns of interactions, usage and other visible indicators (Ellison, 2007). These patterns help researchers to understand how people feel, form and share opinion about other people, institutions, companies, products and issues. When specifically targeted on an institution or person or an issue, this could give us a snapshot of people's perception about the entity or the issue at a point in time. For example, Twitter profile and linkage data could be investigated to understand network's internal structure and dynamics of a news cycle.

Social network theory states that there are a few people in the social network that connect everyone else together. The famous Milgram's small-world experiment (Milgram, 1967), in reality, means that a 'few people' are connected to everyone in a few steps and the rest of them are connected to the world through those 'few people'. People in the network work as connectors who spread the idea, as databanks who provide the message and as salesmen who sway those who are not convinced at what they are hearing (Bruck, 2011). In organizations, these networks often facilitate participatory decision-making (Hashim et al., 2010). Social networks have been used to study the complex set of relationships at micro

*Corresponding author. Email: sameerkr03@yahoo.com.

(individual), meso (local) and macro (global) levels. Social analysts now reason and study whole networks, egocentric networks and less-bounded social systems. Social network analysis (SNA), a method used in the analysis of social networks, focuses on how the structure of ties affects the nodes and their relationships (Borgatti et al., 2009).

Political activists, companies and researchers now believe that activities on social media by millions of people, represent social interactions which could be utilized in the study of propagation of ideas, social bond dynamics and viral marketing, among others (Huberman et al., 2009). A message that arouses the interest of user to be tempted to have it sent along (virality) is the subject of intense study. A recent study found that both negative and positive messages have their share of virality when it comes to news and non-news segment respectively (Hansen et al., 2011). Another study found mindful adoption, community building and absorptive capacity as three important elements for gaining full business value from social media (Culnan et al., 2010). Cheung et al., (2011) found that social related factors had the most significant impact on the intention to use sites like Facebook, by students. Other studies too have found social connection to be an important reason why users use Facebook (Joinson, 2008; Ledbetter et al., 2011). There have been quite a few interesting network based studies on Twitter. We discuss some here. One study, involving 1,348,543 posts from 76,177 unique users, found that people use Twitter primarily to discuss their daily activities and to seek or share information (Java et al., 2007). Huberman et al. (2009) found that people interact with only a small percentage of their "declared" list of friends. Scarcity of time and attention makes people restrict exchange of information with those who matter. Kwak et al. (2010), by crawling the entire Twittersphere, found non-power law "follower" distribution and other traits in its "follower-following" relationship topological analysis, which marked a deviation from the known characteristics of human social networks.

The literature reveals that, although there have been researches conducted to understand how Twitter have been impacting our lives and business, there have been very few studies (Java et al., 2007; Kwak et al., 2010; Huberman et al., 2009) that have actually apply formal methods of SNA in their investigations. Those researches that have applied SNA for the analysis of Twitter have done so for large networks, which fail to illustrate finer interplays of nodes.

In this study, as a case study, an in-depth investigation was carried out on the social structure, dynamics and geo-spatiality of a network formed by those tweeting (a 'tweet' is another name for short message on Twitter) "infosys". Infosys (Infosys Technologies Limited) is India's second largest information technology (IT) Company in terms of turnover (Revenue approx US$5 billion in FY 2011) with an employee base of 1,033,560 and offices in 33 countries. It is active on Twitter with username "Infosys". Infosys have been recently in the news because of top-level changes in the management, resignation of few top executives and eagerness in the stock market with regards to its quarterly financial results. This had generated a lot of talk in the media and authors presumed that it would lead to capturing of more tweets on Infosys for data analysis.

Network of users, formed by virtue of their connections through tweets, was drawn and then a comprehensive social network analysis to understand the structure of network and the communities they form was carried out. The evolution of the network is captured over a 24 h period, which also serves as a demonstration of how even short-duration networks form online.

Rest of the article is organized as follows: Research methods, data-set and tools deployed are discussed next. In the third part, an in-depth analysis of the dataset is carried out, revealing its global and local features, network's evolution, geo-spatiality and formation of communities around tweets. Finally, conclusions are made.

## MATERIALS AND METHODS

### Data mining and analysis tools

On 21st April 2011, real-time data was harvested from Twitter.com's search network extracting tweets that contained keyword 'infosys'. Data was then analyzed by applying social network analysis methods using NodeXL. NodeXL is an open source network analysis and visualization template of Microsoft Excel (Kumar, 2011; Smith et al., 2009). The software uses Stanford network analysis platform (SNAP) for calculating some of its graph metrics.

### Construction of network

Twitter users send a 140-word or less message via computer or mobile device, which is broadcast to those who are 'following' the user.

Twitter offers a very simple interface for interacting with other Twitter users. Users just enter a message in their text box and click "update". This message or tweet is added to the twitter database and will appear on the user's homepage, on user's profile page and on the timeline pages of the people that follow the user. Public timeline database is the list of tweets that all users have been putting go twitter. Relevant tweets can be extracted from this public database using relevant search terms or *hashtags*.

This research considered nodes (the dots in the network) as Twitter users and the link (lines in the network) as the 'following' or 'mentions' relationship on Twitter (Figure 1).

### Twitter common terms

*"follows" and "following"*: Twitter users are connected with others through two types of relationship – "*following*" and "*follows*". These two relationships create three groups that make up the Twitter community – first, those following a user, second, those who the user follows and the third, who both follow and are followed by one

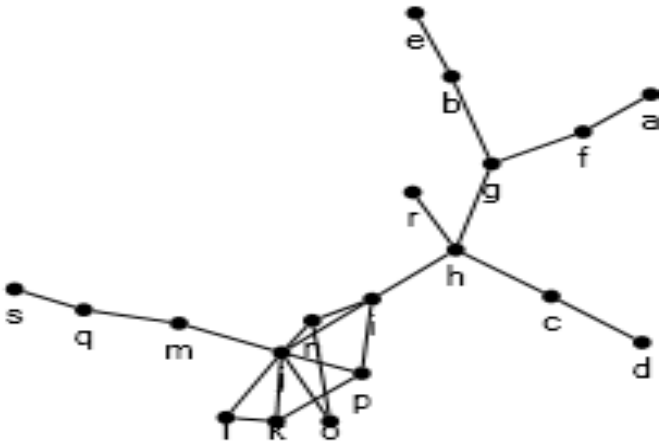



**Figure 1.** An example of a Twitter network where nodes are the individuals and the relationship is either "following" or "mentions".

another. *RT*: means a Re-Tweet. A message sent by a user is re-broadcast by the follower to those who are following the user (that is, RT @sidin). @ - means that a person is specifically addressing his tweet to a particular user (that is, @sidin).

*#hashtag* – tagging word that facilitates in specific search about the topic (that is, #JapanHelp).

*"mentions"* relationship takes place when a twitter user @replies to a specific user or Retweets (RT) to a specific user. User X's tweet may be 'RT'ed or "mention"ed by other nodes in the network, however, User X would become part of the network only when he formally makes a tweet.

### Methods

SNA is used as the method to study the structure and dynamics of our Twitter network. SNA uses mathematical algorithms to analyze the social structure of the network (Wasserman and Faust, 1994). Network diagram is drawn and computation of the graph metrics, clusters and components of the said network, is carried out. Graph metrics are studied at two levels – the global level and the local level. At the global level, the structure of the entire network is taken into consideration. Calculation of measures such as components, clusters (community structure), geodesic distance, diameter, degree distribution, density and the average clustering coefficient, are carried out. For a symmetric graph G with N nodes, density D defined as $D = \frac{2*(\#L(G))}{N(N-1)}$, density of a graph represents the number of edges in the network in ratio to the maximum edges possible.

Clustering coefficient of a network is the average of clustering coefficient of individual nodes in the network. Defined as $C = \frac{3 \times no.of\ triangles}{no.of\ connected\ triples}$, where triangles represent trios of vertices in which each vertex is connected by both others, and connected triples represent trios of vertices in which at least one vertex is connected to both others (Albert and Barabasi, 2002). Community structure is another common characteristic of real-world networks. Communities are formed when there is fewer numbers of connections between two clumps of nodes. There are known algorithms to detect community structure of the network (Girvan and Newman, 2002).

Local level metrics primarily refers to centrality measures (Freeman, 1979) – degree, betweenness centrality, closeness centrality, eigenvector centrality and PageRank. Degree of a node is the number of nodes directly connected to it. Degree centrality, $k_i$ of node *i* is defined as $k_i = \sum_{j=1}^{n} g_{ij}$ , where $g_{ij}$ = 1 if there is link between vertices *i* and *j* and $g_{ij}$ = 0 if there is no such connection.

A variant of degree centrality, eigenvector centrality measure assigns higher values to those nodes that are connected by high-ranking nodes (that is, higher degree of the node) (Lohmann et al., 2010). Denoting eigenvector centrality of node *i* by $x_i$, then by making $x_i$ proportional to the mean of centralities of *i*'s neighbours: $x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j$ , $\lambda$ is a constant and $A_{ij} = 1$ implies the presence of an edge from *j* to *i* (Newman, 2007). PageRank centrality, a variant of eigenvector centrality is an importance measure that uses the Google's PageRank algorithm to assign values to the nodes (Newman, 2007).

Betweenness centrality of node *i* is the fraction of geodesic paths which pass through *i*. Mathematically, betweenness centrality, *b* of node *i* is expressed as $b(i) = \sum_{j,k} \frac{m_{jik}}{m_{jk}}$ , where $m_{jk}$ is the number of geodesic paths from vertex *j* to vertex *k(j, k ≠ i)* and $m_{jik}$ is the number of geodesic paths from vertex *j* to vertex *k*, passing through vertex *i* (Otte and Rousseau, 2002).

Closeness centrality is node's measure of its geodesic distance from all other nodes in the network. Mathematically, closeness centrality $c_i$ of node *i* is written as $c_i = \sum_j d_{ij}$, where $d_{ij}$ is the number of edges in the geodesic path from vertex *i* to vertex *j* (Otte and Rousseau, 2002).

To extract influential nodes, individual nodes were ranked based on the average of their centralities. Applying temporal analysis, growth of network was tracked in the function of time. Network was then laid on geographical map using geospatial co-ordinates of the origin of user. Community structure was detected and frequency analysis of common 'tweet' words each community exchanged between other members were analyzed to distinguish those tweets that were instrumental in the formation of the group.

## RESULTS AND DISCUSSION

### Calculation of global metrics

Two nodes are connected in the network if there is either a "mentions" or "following" relationship. In the network, 92 users had at least one in-degree or out-degree. 84 users formed one giant component, meaning that there is an informal group of users who have tweeted about "infosys" and its nodes interconnect with one another to form one connected clump. 8 other users formed 4 dyadic (component consisting of 2 nodes) components (Figure 2).

The 92 uses form 95 unique relationships with 54 duplicates. Duplicate edge value indicates repeated link between two nodes, signifying the strength of the tie.

### Scale-free characteristics

The in-degree distribution showed (Table 1) certain

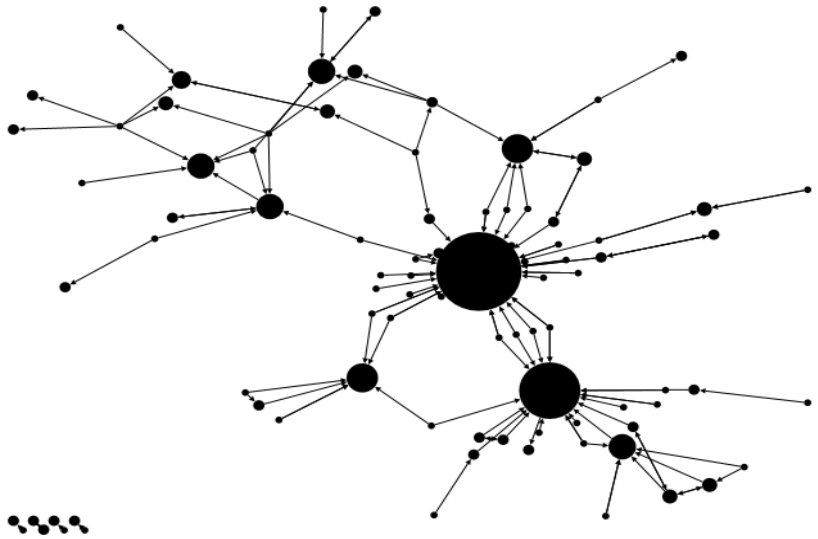


**Figure 2.** Network of users who have tweeted about 'infosys'. Larger nodes size depicts users with more 'tweet' connections.

**Table 1.** In-degree and out-degree matrix.

| In-degree | Out-degree | | | | | |
|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **5** | **Grand Total** |
| 0 | | 28 | 18 | 2 | 2 | 50 |
| 1 | 8 | 10 | 7 | 1 | | 26 |
| 2 | 3 | 1 | 2 | 1 | | 7 |
| 3 | | 1 | | | | 1 |
| 5 | 1 | 2 | 1 | | | 4 |
| 6 | 1 | 1 | | | | 2 |
| 17 | | 1 | | | | 1 |
| 30 | 1 | | | | | 1 |
| Total | 14 | 44 | 28 | 4 | 2 | 92 |

nodes having relatively large number of connections when compared with other nodes in the network (as the concept of degree distribution of a network and degree centrality are related, there would be discussion overlaps). There was one node having 30 connections and another having 17 connections. However, 76 of the 92 connections either had no in-degree or just 1 in-degree. Similar pattern were seen for out-degree too. However, the in-degree distribution seemed to be more skewed (hence suggesting a longer tailed power law) when compared with the out-degree distribution.

The general feature of real world networks is that few nodes have lot of connections and majority of nodes have few or no connections. This feature of the network is known as the scale-free characteristics of a network. One determining feature of scale-free networks is the existence of hubs or popular nodes having relatively larger number of connections. Hubs are created due to nodes propensity to link with the other nodes in the network that are already well connected. However, preference to connect to a certain node could also be the result of some other similarity or dissimilarity in attributes of the nodes, referred to as homophily or assortativity (Newman, 2003). Degree distribution of a network calculates the number of direct connections of each node to determine whether or not the distribution is skewed. Statistics of degree distribution showed (Table 1) that there were certain nodes, which are relatively more "active" than the rest of nodes in the network, suggesting scale-free nature of the network.

### 'Small-world' property

Geodesic distance indicates the level of randomness of a network. Most of the real work networks are somewhere in between the two extremes and the distance between any two randomly chosen nodes in the network is less.

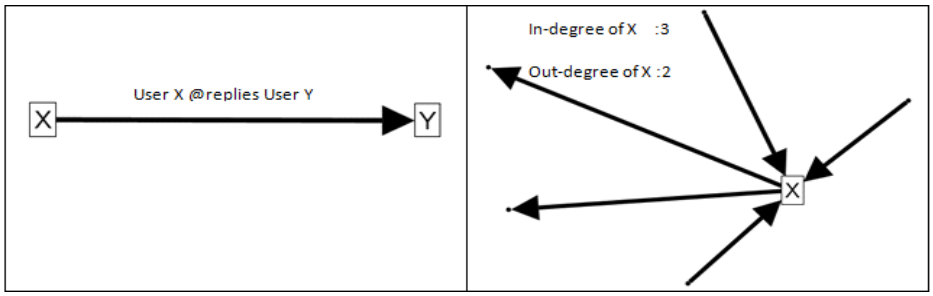



**Figure 3.** Representation of in-degree and out-degree of a node.

Also known as the small-world property, this feature, in most cases, is the result of scale-free characteristics of networks. The diameter of the network was 9 and average geodesic distance was about 4. These values indicate that the network has small-world properties. Maximum vertices and edges in the connected component was 84 and 142 respectively.

### Graph density

The density of the network is 0.0145, suggesting that the network was not dense in nature. A fully connected network would have a density of 1.

### Clustering coefficient of a network

Clustering coefficient is a measure to gauge nodes' propensity to form a clique (in a clique all the nodes are connected with all other nodes), which is also similar to the concept in sociology known as the "fraction of transitive triples" (Wasserman and Faust, 1994). Most of the real world networks also have a high clustering coefficient (Watts and Strogatz, 1998). If the clustering coefficient is 0.5, it means that the chance of nodes to form a clique is fifty percent. The network had the average clustering coefficient of 0.066 which indicates its propensity to form a clique is about 6.6%.

### Calculation of local metrics

Centrality of a node was investigated using the five known centrality measures – degree (in-degree and out-degree), betweenness, closeness, eigenvector and PageRank. Degree, closeness and betweenness centralities reveal the most central or influential nodes in the network, whereas eigenvector and PageRank reveal those that are most prestigious.

### Degree centrality

As the edges formed between the nodes are directed [Figure 4, depicting a "following" relationship, @replies to or RT (Re-tweet)], the in-degree and the out-degree of individual nodes were calculated separately. In-degree of a node is the number of nodes pointing towards it and the out-degree is the number of edges pointing outwards of it. An in-degree edge is drawn when, for example, userX @replies user Y (Figure 3).

Only 39 of the 149 relationships formed because of the tweet that "mentioned" the user in context. The remaining 110 relationships formed because they "followed" the user. There were 50 nodes out of 92 connected nodes that had 0 in-degree. But these same nodes had comparatively very high out-degree. Similarly, 18 nodes with 1 in-degree had 1, 2 or 3 out-degrees (Table 1). One user, "sidin", outsmarted all others in terms of number of connected edges with respect to in-degrees. However, the same node had zero out-degree, meaning that the user was not following anyone who was part of the present network or had specifically replied/RT anyone in the present network.

### Betweenness centrality

Average betweenness centrality of the network was 240.2. User "sidin" had the maximum betweenness centrality of value 5232. "sidin" forms a bridge or a cut-point between two or more network communities in our network. Removal of node "sidin" is likely to break the network into pieces.

### Closeness centrality

The closeness centrality of 2-node components stood at 1.0. These two-node components were detached from the giant component but still command a high closeness centrality. Hence closeness centrality values may be misleading when taking into consideration all components in the network. In the giant component, user "sidin" had the maximum closeness centrality measure of 0.005. Nodes with high closeness centrality tend to occupy position at the centre of the network (of the giant component) when drawn using force-directed graph

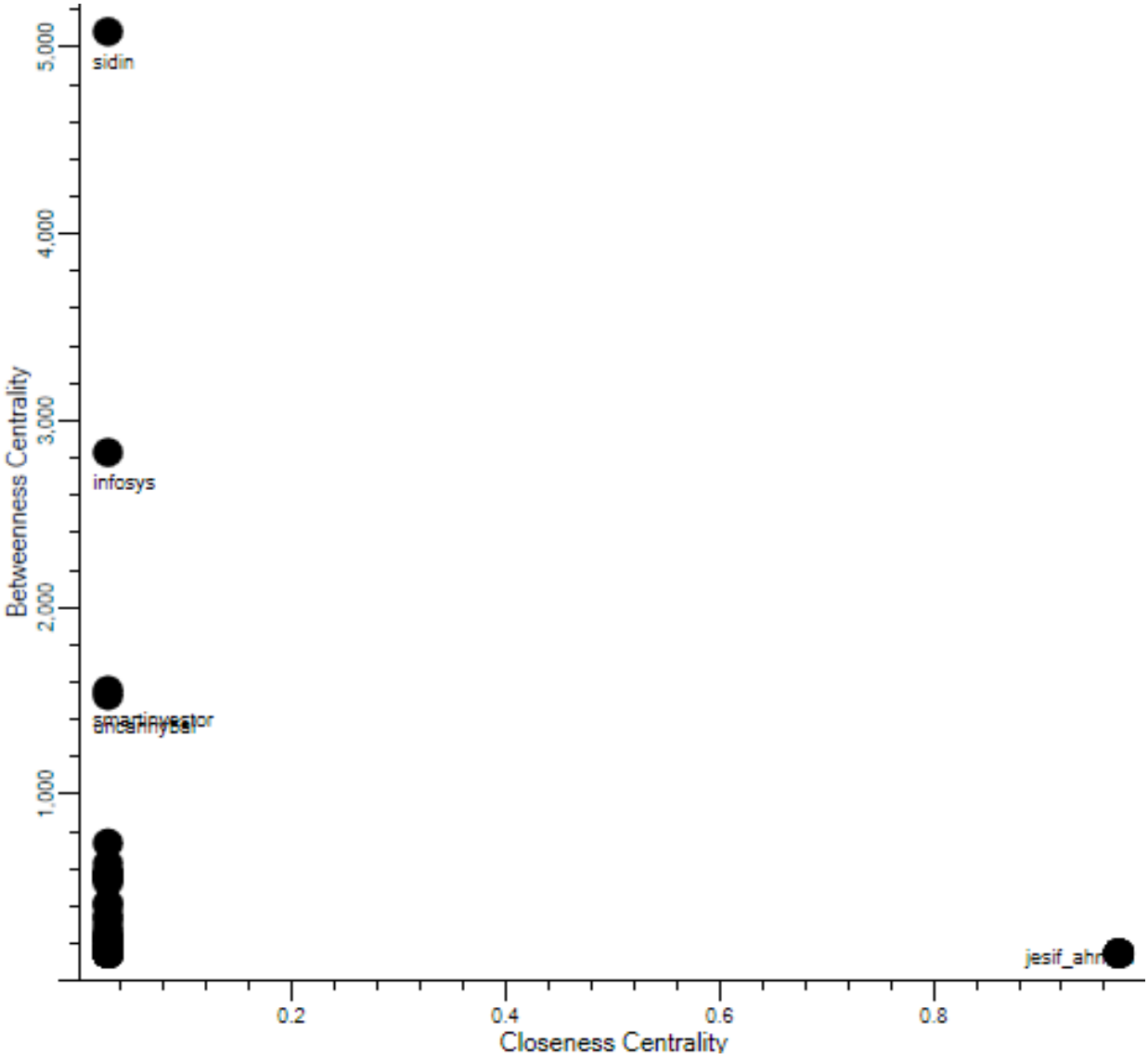



**Figure 4.** XY Graph of betweenness and closeness centrality measures.

layout algorithms such as Fruchterman-Reingold algorithm (Fruchterman et al., 1990).

The two important centrality measures, betweenness and closeness were laid out using an XY graph to detect outliers in our network. The XY graph revealed an unexpected outlier – user "jesif_ahmed" (Figure 4). However, this user was part of a 2-node network component and detached from the largest group (giant component). Apart from 'jesif_ahmed', the graph revealed only 'sidin' as the prominent outlier.

## Eigenvector and PageRank centrality measures

User "sidin" again had the highest eigenvector and PageRank centrality measures of 0.107 and 10.851 respectively, signifying that the user was also receiving 'quality' connections.

## Ranking of nodes

As mentioned earlier, degree, closeness and betweenness centralities reveal the most central nodes, and the eigenvector and PageRank centralities reveal those that are prestigious; hence, a rank was necessary to see the overall importance (those who are both central and prestigious) of nodes. We ranked nodes based on the average of five centrality measures (Degree- in and out, closeness, betweenness, eigenvector and PageRank). Our ranking system found "sidin" as the top ranked user followed by "smartinvestor" "adeobhak" and

**Table 2.** Ranking of nodes based centrality measures.

| Vertex | In-degree | Out-degree | Betweenness centrality | Closeness centrality | Eigen-vector centrality | PageRank | Clustering coefficient | Rank |
|---|---|---|---|---|---|---|---|---|
| Sidin | 30 | 0 | 5232 | 0.004878 | 0.107393 | 10.850709 | 0.00344828 | 1 |
| smartinvestor | 5 | 2 | 1550.7476 | 0.003356 | 0.004355 | 2.127488 | 0.06666667 | 2 |
| adeobhak | 1 | 2 | 0 | 0.003497 | 0.024906 | 0.795541 | 0.5 | 3 |
| apoorvvv | 1 | 2 | 0 | 0.003497 | 0.024906 | 0.795541 | 0.5 | 3 |
| vijaybhargava | 0 | 2 | 563.66667 | 0.004082 | 0.02721 | 0.763946 | 0 | 5 |
| seedhesadheakki | 0 | 2 | 563.66667 | 0.004082 | 0.02721 | 0.763946 | 0 | 5 |
| _bharath | 0 | 2 | 563.66667 | 0.004082 | 0.02721 | 0.763946 | 0 | 5 |
| urmilesh | 0 | 2 | 563.66667 | 0.004082 | 0.02721 | 0.763946 | 0 | 5 |
| anuvratbhansali | 1 | 2 | 164 | 0.003509 | 0.021538 | 0.915751 | 0 | 9 |
| gautamghosh | 6 | 1 | 731.1 | 0.003205 | 0.013045 | 2.136931 | 0 | 10 |

"apoorvvv" (Table 2). In Figure 5, a graph is drawn based on ranking.

## Temporal analysis

Addition of nodes and edges in the network in the function of time reveals the dynamics of network. Understanding how the network evolves and what causes this evolution are accomplished using temporal analysis.

As seen, the network was forming in the early morning (Figure 6a) and then developing as the day went by (Figure 6b, c and d). The tweets till 8:11 AM were about media buzz on the change in top-level management of Infosys. The interest from stock analyst was also seen as they tweet about the stock prices when the Indian markets open. Some light comments were also seen tweeted, reflecting the mood of the day.

The tweets continued when the stock market opened, indicating the interest in the stock market about Infosys and even the slightest move in the markets was tweeted back so that investors can take heed and act accordingly.

The tweets during the evening was about who would take the new position of chairman. Some are still speculating about the sudden resignation announcement of 'Mohandas Pai' as the Director of Infosys. As the day passes, the main topics of the tweets change suggesting the general mindset of people at different times of the day.

## Geospatial analysis

Geographic maps depict spatial dynamics. Geospatial co-ordinates of the tweets were investigated based on the time zone of user's city. City of origin of users was available for 148 tweets of the 200 tweets. For the purpose of this study, Northern America time zones are assumed. In the USA, we used cities such as Hawaii, san Francisco, Phoenix, Houston and New York City as representative cities for the following times zone, Hawaii Time, Pacific Time, Mountain time, Central and Eastern Time, respectively. Latitude and longitude were found for each city by geo-coding using gpsvisualizer.com. The geo-codes, name of the city and frequency of its occurrence, give inputs to draw geospatial diagram (Figures 7).

Majority of the tweets originate from either Indian time zones or from US and Canada time zones. Over 60% of tweets have originated from Indian time zones which prove that domestic interest about the company is high. Infosys has stakes in USA and number of tweets from Northern America confirms this.

In order to have trackpoint for each tweet, a code sheet was created (Table 3) and then the co-ordinates were overlayed on the map. Perspective of relationship brought in by tweets is shown in Figure 8.

## Cluster analysis and evolution of the largest group

Girvan-Newman community detection algorithm (Girvan and Newman, 2002) was used to divide the network 11 communities were detected, each

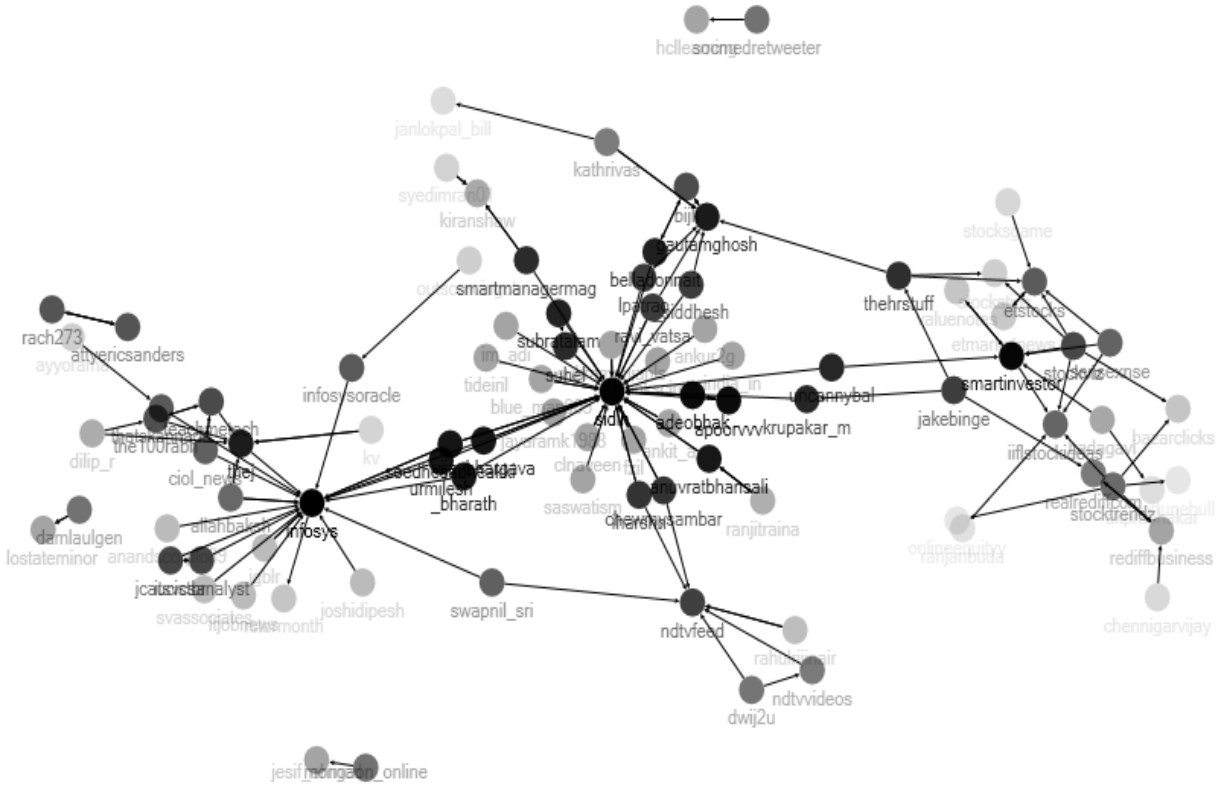



**Figure 5.** Graph based on Rank; darker circles indicate higher ranked nodes.

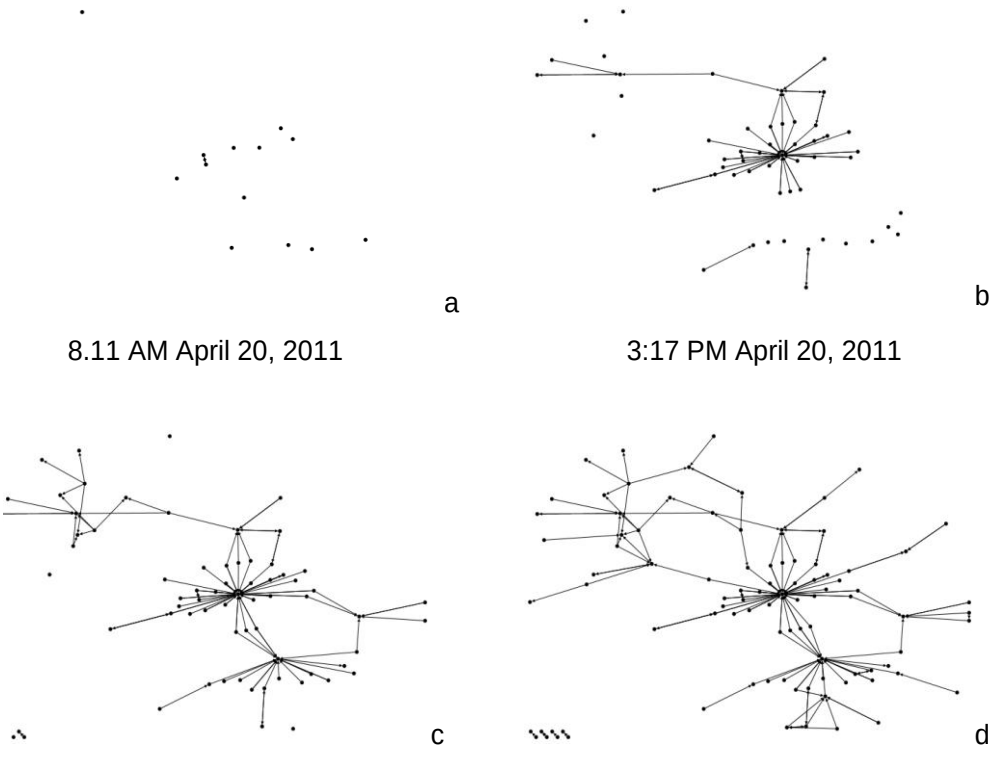



**Figure 6.** Evolution of twitter network.

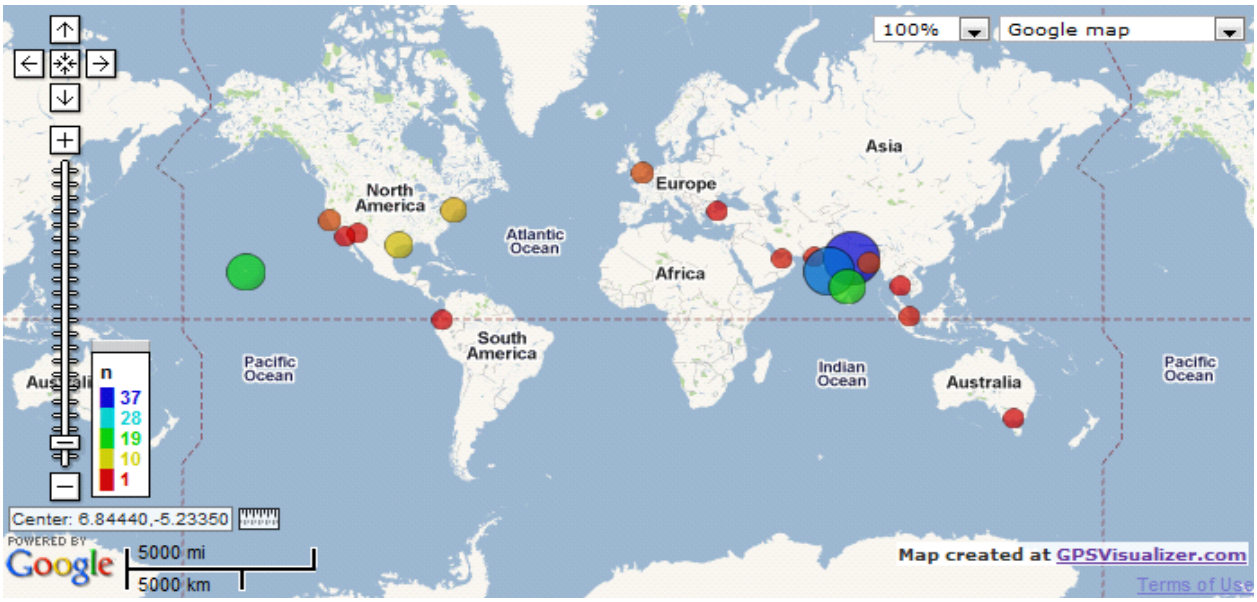



**Figure 7.** Tweets emergence; bigger circles indicate larger number of tweets from that region; big green dot on the left in the middle of the Pacific Ocean indicates Hawaiian Islands.

**Table 3.** Showing top 14 trackpoints overlayed on the actual world map (Google).

| Name | Desc | Latitude | Longitude | Relationship | Relationship date (UTC) |
|---|---|---|---|---|---|
| im_adi | Mumbai | 19.076191 | 72.875877 | Mentions | 4/20/2011 8:34 |
| Sidin | London | 51.506325 | -0.127144 | | |
| joshidipesh | New York City | 40.71455 | -74.007124 | Followed | 4/21/2011 5:09 |
| Infosys | New Delhi | 24.7338 | 81.33463 | | |
| svassociates | Mumbai | 19.076191 | 72.875877 | Followed | 4/21/2011 5:09 |
| infosys | New Delhi | 24.7338 | 81.33463 | | |
| anandscorpio89 | Chennai | 13.06397 | 80.24311 | Followed | 4/21/2011 5:09 |
| Infosys | New Delhi | 24.7338 | 81.33463 | | |
| clnaveen | Chennai | 13.06397 | 80.24311 | Mentions | 4/20/2011 8:07 |
| Sidin | London | 51.506325 | -0.127144 | | |
| clnaveen | Chennai | 13.06397 | 80.24311 | Followed | 4/21/2011 5:09 |
| Sidin | London | 51.506325 | -0.127144 | | |
| Fzil | Chennai | 13.06397 | 80.24311 | Mentions | 4/20/2011 8:08 |
| Sidin | London | 51.506325 | -0.127144 | | |

containing a certain number of nodes. G1 with 27 and G2 and G3 with 18 and 14 nodes respectively are most significant groups in the network. Groups with just 2 or 3 nodes are insignificant as the size of their group is still comparatively small. G1 is the most active group; however, it received only in-degree links from other groups. G2 received both in-degree and out-degree links. Only G3 has a mutual link (Figure 9)

## Evolution of the largest group

Group G1 is the largest group in the network with 27 nodes. User “sidin” has the maximum degree and the maximum betweenness. The star of G1, ‘sidin’ may as well be called the star of the whole network.

A closer analysis of the formation of G1 (Figure 10) revealed that, most of the network was formed due to retweets of user “sidin” tweet posted before the capture of data. The first tweet of user “sidin” got re-tweeted several times and when ‘sidin’ came to the scene with another tweet, the nodes which had already re-tweeted sidin’s earlier tweet, got connected with “sidin”. The results also indicate that a large proportion of communications is either through @replies or RT. RTs here seem to be playing a role in the diffusion of

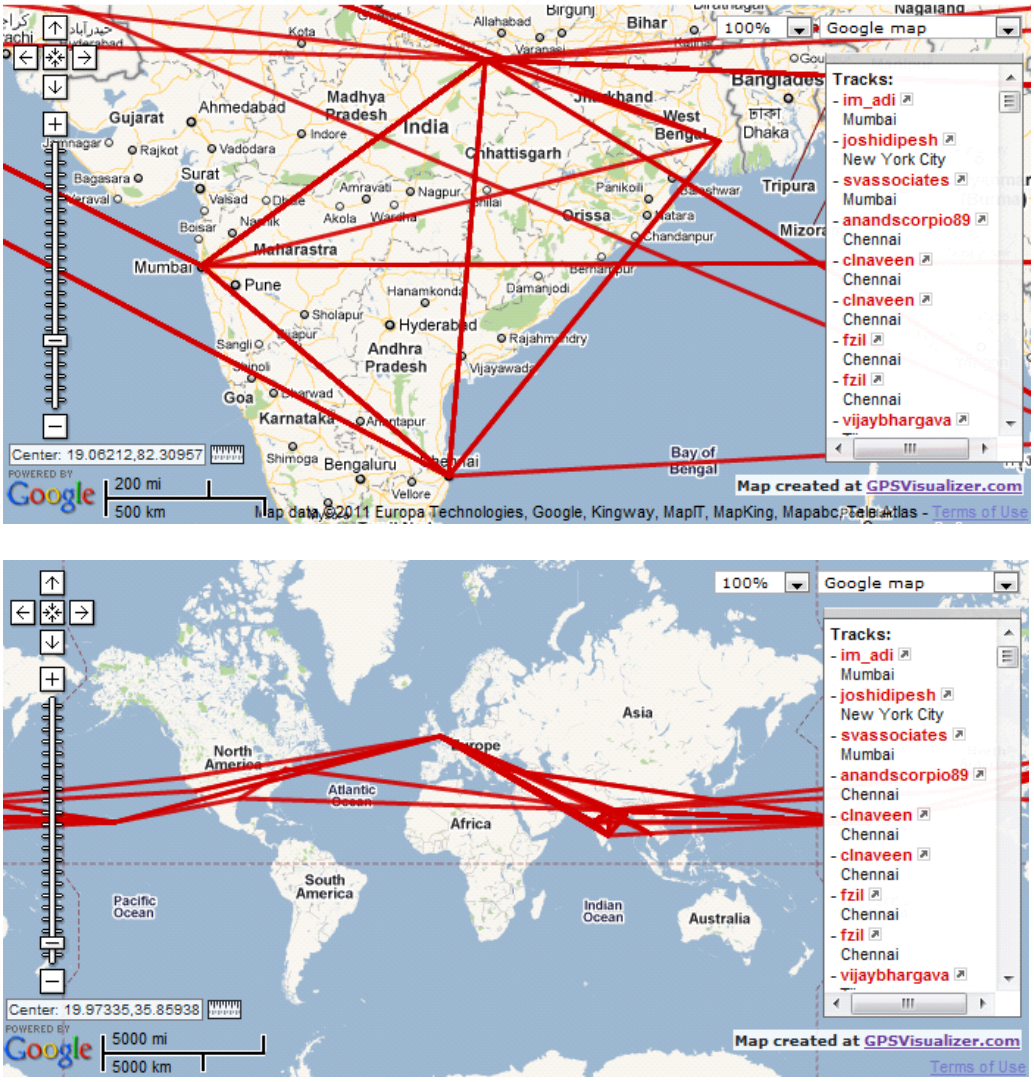



**Figure 8.** A tweet network on geographical map (unweighted).

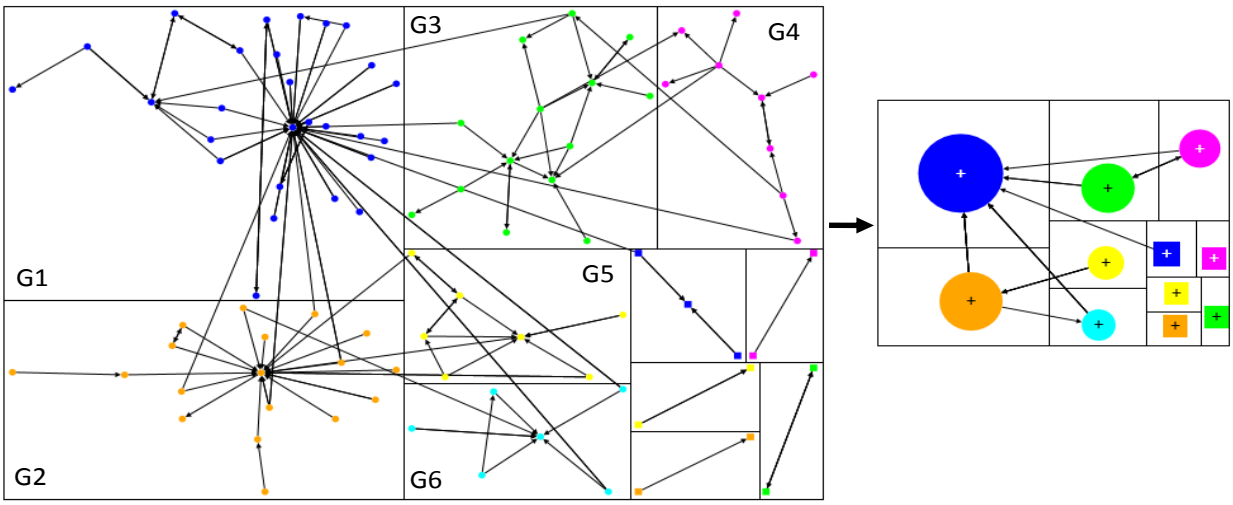
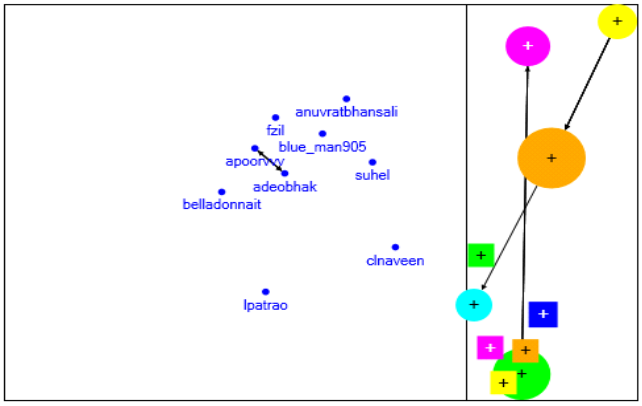
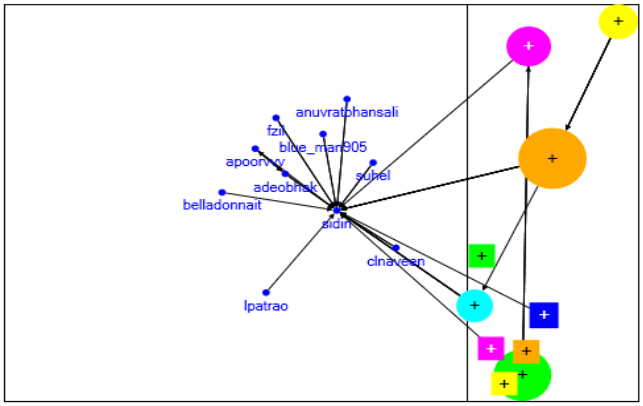


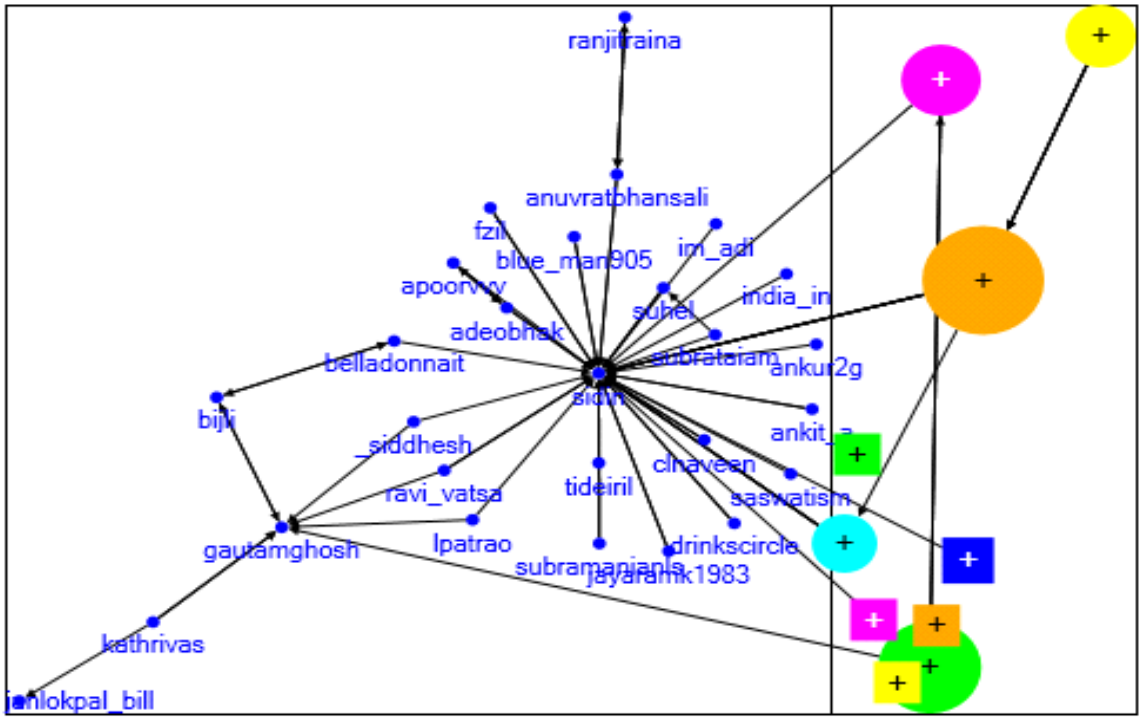

**Figure 9.** Network divided into sections using Girvan-Newman algorithm.




**TRANSITION 1:** G1 Group formation at 8:27 AM (April 20th)




**TRANSITION 2:** G1 Group formation at 8:29 AM (April 20th)




**TRANSITION 3:** G1 Group formation at 4:23 AM (April 21th)

**Figure 10.** Evolution of the largest group.

information. RTs also seem to be instrumental in the formation of the network itself. It was found that there were less initiator to the tweet, than people who RT and @reply to the first initiated tweet.

### Frequency analysis of common words in tweets

Detailed analysis was carried out for each tweet that was tweeted within the group. Such an analysis helps in the understanding what "talk" actually went on in the group that led to the formation of the group. Frequency calculation of each word tweeted within the group was done and then frequent and meaningful words were skimmed out (Table 4).

In the largest group, G1, the most frequent words were "sidin", "RT" ,"CEO", "coach", "India" and "team". Content analysis of group evolution revealed that this group was formed primarily by just two retweets (RT) of "sidin":

1) RT @sidin: If you could choose between becoming

**Table 4.** Frequency analysis of common words in tweets.

| Group Words | G1 | G2 | G3 | G4 | G5 | G6 | Grand total |
|---|---|---|---|---|---|---|---|
| Infosys | 27 | 22 | 17 | 13 | 10 | 7 | 96 |
| http | 11 | 12 | 11 | 8 | 3 | 5 | 50 |
| RT | 15 | 3 | 1 | 1 | | 2 | 22 |
| CEO | 12 | 2 | 2 | 1 | 1 | 2 | 20 |
| Sidin | 16 | 2 | | | | 2 | 20 |
| Mohandas | 2 | 4 | 2 | 7 | 1 | | 16 |
| Pai | 2 | 3 | 2 | 7 | 1 | 1 | 16 |
| News | 6 | | 2 | | 3 | | 11 |
| Coach | 8 | 2 | | | | | 10 |
| India | 9 | | | | | | 9 |
| Team | 9 | | | | | | 9 |
| Buy | | | 5 | 3 | | | 8 |
| Livemint | 6 | | | | | | 6 |
| Discord | 3 | | 1 | | | | 4 |
| Results | | | 4 | | | | 4 |
| Clients | | | | | | 3 | 3 |
| Confidence | | | | | | 3 | 3 |
| Dip | | | | | | 3 | 3 |
| Equities | | | | 3 | | | 3 |
| Generations | | | | 3 | | | 3 |
| Job | | 3 | | | | | 3 |
| Kamath | | | 2 | | | 1 | 3 |
| Kris | | | 2 | | | 1 | 3 |
| Leaders | | | | 3 | | | 3 |
| Lost | | | | 3 | | | 3 |
| Quit | | | | 3 | | | 3 |
| Leader | | | | | 2 | | 2 |
| Reports | | | | | 2 | | 2 |
| Stocks | | | 2 | | | | 2 |
| TCS | | 2 | | | | | 2 |
| Shibulal | | | | | | 1 | 1 |

coach of Team India and CEO of Infosys... which one would you choose?
2) RT @sidin: Inheriting Infosys - corporate news - livemint.com http://bit.ly/eme9pP.

The discussion was around the recent media buildup on Mohandas Pai and speculations about the next CEO of Infosys. This was evident from the words, “Mohandas”, “Pai” and “CEO” in all the groups (G1 to G6). In G2 there was also a mention of “TCS”, a company whose quarterly results were recently announced. In G3 the talk was about “stocks”, “buy” and “results”, indicating there were some stock advisors active in the group who were keeping a close tab the financial news about Infosys. In G3 there was a mention of “Kamath” and “Kris”. Kamath was tipped to be the next chairman of Infosys, replacing the incumbent Narayana Murthy, and this discussion also surfaces. Kris is implied for K. Gopalakrishnan (co-founder Infosys) who, as per news in the media, was expected to take charge as the CEO. In G6 there was mention of “shibulal”. Shibulal is also a co-founder of Infosys and there was a rumour that he would be the next CEO.

It is seen that prominent words form the core of the network and those with less group-wise frequency form the periphery of the network. The thickness of the edges indicates the frequency of words within the group.

A unique 2-mode network diagram is drawn to illustrate the tweets in groups, how they are held together within the group and how the words inter-relate with same words from other groups (Figure 11).

Frequency of word “infosys” is irrelevant here as the whole network extraction has taken place on this keyword. Across the top 6 groups, “http” is the most common co-word, indicating that most tweets also refer to an address (http) on the WWW. Majority of these webs

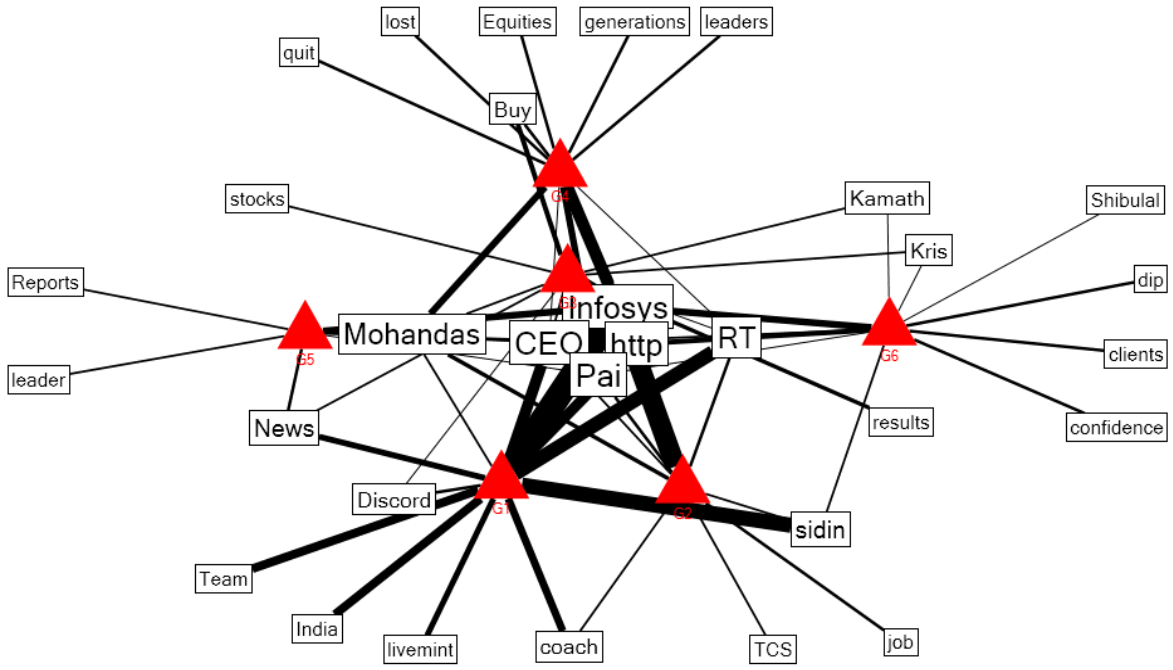



**Figure 11.** A two-mode graph showing groups (red triangles) and words (textboxes).

addresses point to a newsmedia article or website, indicating that people tweet what they hear in newsmedia.

## Conclusions

The purpose of this study was to understand the structure and evolution of a Twitter network. Real-time data was harvested from Twitter's search network of those tweeting about a company with keyword "Infosys". In-depth analysis of this small 200-node Twitter network formed during a 24 h period was carried out to reveal its structure and dynamics. Five centrality measures – degree, betweenness, closeness, eigenvector and PageRank, were computed and assigned ranks to individual nodes based on the average of the five centrality measures. Community structure was detected and frequency analysis of common words each community is exchanging was analyzed to detect those tweets that were instrumental in the formation of the group. The study revealed that the Twitter network had both the small-world and scale-free characteristics. Computation of centralities and global metrics revealed the overall stars of the network that were both influential and prestigious. Temporal analysis revealed the network dynamics in the function of time. The study found an increased network formation activity when there was a tweet of interest. Specific communities were formed in the network based on tweets of similar theme. The network also had a common theme that held the communities together. A unique 2-mode graph was drawn to demonstrate how words held together within the group and inter-linked with one another in the network. Geospatiality revealed that tweets mostly originate from regions where the subject is physically located or has specific business or other interests. The central node in the network was the one whose tweets were replied or re-tweeted ('preferential attachment' or homophily) the most. Conversely, the study found that it was not necessary for the central node (or nodes) to be doing the same (that is, retweeting) for others' tweets. Retweets (RT) and @replies are an important mechanism for the trigger of more tweets and hence a prominent cause of network formation. The user whose tweets were retweeted the maximum ('sidin' in our case) were also likely to have larger centrality (or influence and prestige) points. These nodes also keep the network together and are instrumental in the formation of the group. Tweets were generally about the topic, which are in the newsmedia, indicating that news is harbingers to tweets on Twitter.

SNA have been applied in a wide variety of settings - from anthropology to neural networks. This is probably one of the first applications of SNA that investigates the structural properties of a small information network formed on Twitter in a 24 h period. It would be interesting to see whether networks formed around non-firm tweet keywords - for example, disaster management (that is. JapanHelp), people's movement (that is, Egypt), etc., display similar structural properties and patterns as found in the present research.

By doing a micro study to understand the topological properties of a small network formed on Twitter we add to the growing body of Twitter literature, hitherto dominated by macro studies of large Twitter networks.

## ACKNOWLEDGMENT

Authors would like to acknowledge research support provided by Asia-Europe Institute, University of Malaya, toward this end.

### REFERENCES

Albert R, Barabasi AL (2002). Statistical mechanics of complex networks. Rev. Mod. Phys., 74(1): 47-97.

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009). Network Analysis in the Social Sciences. Science, 323(5916): 892-895.

Bruck J (2011). Designing for Social Networks. Retrieved from http://www.jonbruck.com/thoughts/papers/design_socialNetworks.pdf.

Chen GM (2011). Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. Comput. Hum. Behav., 27(2): 755-762.

Cheung CMK, Chiu PY, Lee MKO (2011). Online social networks: Why do students use facebook? Comput. Hum. Behav., 27(4): 1337-1343.

Chew C, Eysenbach G (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. Plos One, 5(11).

Culnan MJ, McHugh PJ, Zubillaga JI (2010). How Large U.S. Companies Can Use Twitter And Other Social Media To Gain Business Value. Mis Quart. Exec., 9(4): 243-259.

Ellison NB (2007). Social network sites: Definition, history, and scholarship. J. Computer-Mediated Commun., 13(1): 210-230.

Freeman L (1979). Centrality in social networks: I, conceptual clarification. Soc. Netw., pp. 215-239.

Fruchterman TMJ, Reingold EM, Science U (1990). Graph drawing by force-directed placement. Softw.Pract. Experience, 21: 1129-1164.

Girvan M, Newman MEJ (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A., 99(12): 7821-7826.

Hansen LK, Arvidsson A, Nielsen FÅ, Colleoni E, Etter M (2011). Good Friends, Bad News-Affect and Virality in Twitter. CoRR abs/1101.0510.

Hashim F, Alam GM, Siraj S (2010). Information and communication technology for participatory based decision-making-E-management for administrative efficiency in Higher Education. Int. J. Phys. Sci., 5(4): 383-392.

Huberman BA, Romero DM, Wu F (2009). Social networks that matter: Twitter under the microscope. First Monday, 14(1): 8.

Java A, Song X, Finin T, Tseng B (2007). Why we twitter: Understanding micro blogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis. ACM.

Johnson KA (2011). The effect of Twitter posts on students' perceptions of instructor credibility. Learn. Media Technol., 36(1): 21-38.

Joinson AN (2008). Looking at, looking up or keeping up with people?: Motives and use of facebook. Paper presented at the Conference on Human Factors in Computing Systems (CHI), Florence, Italy.

Kumar S (2011). Analyzing social media networks with NodeXL: insights from a connected world. Inf. Res. Int. Electronic J., 16(2).

Kwak H, Lee C, Park H, Moon S (2010). What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web.

Ledbetter AM, Mazer JP, DeGroot, JM, Meyer KR, Mao Y, Swafford B (2011). Attitudes toward online social connection and self-disclosure as predictors of facebook communication and relational closeness. Commun. Res., 38(1): 27.

Lohmann G, Margulies DS, Horstmann A, Pleger B, Lepsien J, Goldhahn D, Schloegl H, Stumvoll M, Villringer A, Turner R (2010). Eigenvector centrality mapping for analyzing connectivity patterns in FMRI data of the human brain. PloS One, 5(4): e10232.

Milgram S (1967). The Small World Problem. Psychology Today, 1: 61-67.

Newman MEJ (2003). The structure and function of complex networks. SIAM Rev., 45(2): 167-256.

Newman MEJ (2007). The mathematics of networks. The new palgrave encyclopedia of economics. 2nd edition, Palgrave Macmillan, Basingstoke.

Otte E, Rousseau R (2002). Social network analysis: A powerful strategy, also for the information sciences. J. Inf. Sci., 28(6): 441-453.

Smith MA, Shneiderman B, Milic-Frayling N, Mendes Rodrigues E, Barash V, Dunne C, Capone T, Perer A, Gleave E (2009). Analyzing (social media) networks with NodeXL. Proceedings of the Fourth International Conference on Communities and Technologies - CandT. p. 255.

Wasserman S, Faust K (1994). Social Network Analysis, Methods and Applications (First edition ed.): Cambridge University Press.

Watts DJ, Strogatz SH (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684): 440-442.

Ye SZ, Wu SF (2010). Measuring Message Propagation and Social Influence on Twitter.com. In L. Bolc, M. Makowski A, Wierzbicki (Eds.), Soc. Inform., 6430: 216-231.