IS482 Analytics Practicum

Kolaveri Di Social Analytics Project

Midterm Report

Submitted on 15th September, 2014

Prepared by:

Chan Wei Yin

Lee Jaehyun

Introduction

The use of social media, such as Facebook, Twitter, or Instagram, among the Internet users has been exponentially increasing in the recent years. The users in social media platforms form "networks" whereby they are connected with one another directly or indirectly. The insights that these social networks give are very valuable in different areas of study, especially in marketing.

For this project, we will be focusing on a video that went viral on social media platforms, especially among the Twitter users. The project will be mainly focusing on a Tamil song, entitled *Why This Kolaveri Di* (Why this murderous rage, girl?) -- a music video that went viral in 2011 on social networking sites for its catchy Tanglish (Tamil and English) lyrics. We aim to evaluate the Twitter dataset related to Kolaveri Di, in order to find out *why and how* the video went viral and to suggest the implications of such phenomenon with marketing insights. We aim to do extensive research to understand the dataset and do hands-on analytics using different tools such as NodeXL.

The Virality of the Video

Achievements (focused on Twitter and YouTube)

The video was officially uploaded on YouTube on 16th November, 2011. From 16th November 2011 to 5th December 2011, it generated a total number of 96,323 tweets with the hashtag #whythiskolaveridi, and the Twitter platform garnered over 8 million impressions.

The video had amassed over 16.5 million views on YouTube (on Official Sony Music Channel), of which 11 million views came from India. As a result, Sony was able to gain nearly 8,500 new subscribers on Youtube. The video also received a number of positive feedback from the viewers: not only did it receive 146,224 likes but also a gold medal for the most popular video and silver medal for trending.[1]

Objectives of this Project

The main objective of this project is to find out *why* and *how* the video went viral. To further supplement, below is the list of questions to be explored throughout the project:

1. **Why did the video go viral?**
    a. What are the factors that made *Kolaveri Di* go viral?
    b. What are the other general factors (or *triggers)* that can make any content go viral on social media sites?
    c. Are there any networks formed among the users? What are some characteristics of the networks formed?

[1] Menezes, J. (December 9, 2011). Why this Kolaveri a rage-u on social media? *Social Samosa.* Retrieved September 14, 2014 from http://www.socialsamosa.com/2011/12/why-this-kolaveri-a-rage-u-on-social-media/

2. **Who are the influencers?**
    a. Define the influencers in both marketing and analytics perspective.
        i. What are the attributes that an influencer should possess?
    b. Identify the influencers in the *Kolaveri Di* network.
        i. What are their characteristics like?
        ii. Profile them.
        iii. Are the influencers the same in different time periods?

3. **What implications does this project have?**
    a) How do we engage/incentivize the influencers in social media marketing?
    b) Make any other feasible recommendations for another video to be successful/go viral in Twitter.

Why Did *Kolaveri Di* Go Viral?

A few factors have contributed to the virality of the video Kolaveri Di. In this project, we will be looking into mainly three reasons:

1. The marketing activities done by Sony and the societal factors
2. The content of Kolaveri Di
3. The networks formed around the influencers and the users in Twitter

1. The Marketing Activities Done by Sony and Societal Factors

When the video was illegally leaked on YouTube 2 weeks before the official release, Sony decided to take the leak as an opportunity and tapped on it, instead of taking the video down. The social connection of Sony played a big part in the spread of the video. Sony had more than 200,000 followers on Facebook, which was a good ground to reach out to the audience. Heavy marketing activities were done by Sony on its social media platforms; it posted the official song on Tamil, Hindi, and International Facebook pages to drive interest and engagement. As a result, the twitter mentions rose by 200% everyday.[2]

However, what comes as more surprises is that *Kolaveri Di* was being talked about among the social media users even before the launch of the offical video, perhaps due to the leak. The hashtag #whythiskolaveridi was created in Twitter by Sony and used in various tweets to generate hype and interest for the official launch of the video. Tweets such as "Country X has not heard of #whythiskolaveridi" or "Not called for 2 days #whythiskolaveridi" were posted, which intrigued people who did not understand what it meant. After the video was launched, people clicked on the YouTube link, as a natural impulse to quench their curiosity.[1]

The virality of the video is also supported by the better social connectivity among the Indian internet and mobile users. India in the year 2011 experienced phenomenal changes in its digital landscape, including better internet connection, e-commerce, or social networks. There were nearly 80 million internet users, and 10 million subscribers on 3G services on their mobile. The Indian internet users quickly became very active on

[2] MSLGroup. (2012). Why Kolaveri Di went viral – Lesson for marketers. *Slideshare.* Retrieved October 1, 2014 from http://www.slideshare.net/mslgroup/rhythm-correct

social networking sites. By August 2011, India had over 33 million registered users on Facebook, which is the 3rd highest in the world. Twitter became one of the most preferred mico-blogging site, with 3 million users including celebrities and business executives.[3]

2. The Content of Kolaveri Di

Based on Kolaveri Di and other videos that went viral online, what are some factors that actually make anything go viral on social media? And how were they applied in Kolaveri Di video? According to Archer (2013)[4], there are six main factors that make a video go viral online. The table below summarizes how the video Kolaveri Di portrayed these factors:

No.	Factors	Description	How was it shown in *Kolaveri Di*?
1	Emotion	The content should be able to build some emotional connection with the viewers and provoke emotions from the viewer (e.g. fun, anger, sadness)	The video conveyed a feeling of heartbreak, but it was still fun to watch.
2	Surprise	The content should not be predictable; if so, the viewers would easily "bounce" out from the content without finishing it	The video was different from other Indian videos or bollywood films, wherein a lot of clichés are used.
3	Intensity	It is important to grab the audience's attention from the beginning, and to keep it through brevity and density.	The non-sensical lyrics, the foot-tapping kind of beat.. it was simple but intriguing and hooked the people
4	Relevance	The content should be relevant to the target audience. Thinking in their perspectives is more important than focusing solely on the messages.	"Is there anyone who hasn't experienced a heartbreak?" People would connect better to the guy in *Kolaveri Di*, than other handsome actors shown in movies.
5	Validation	People tend to share things that support their own perspectives, and that represent their beliefs and opinions.	In fact, 72% of the viewers were male, as they found the content more relevant and connecting to them.
6	Style	The style of how the content is presented is important. Is it presentable? Funny? Engaging?	The video was kept simple and stupid, catchy and fun.

[3] Khedekar, N. (December 27, 2011). Trends of 2011 – digital goes mainstream. *Tech 2.* Retrieved September 21, 2014 from http://tech.firstpost.com/news-analysis/trends-of-2011-digital-goes-mainstream-24297.html

[4] Archer, J. (May 3, 2011). Why some videos go viral. *Inc.* Retrieved September 21, 2014 from http://www.inc.com/james-archer/why-some-videos-go-viral.html

3. The Networks Formed in Twitter

Due to the nature of social media, the networks among the users could be quickly formed which triggered the spread the video faster than expected. In Twitter, the networks will begin with one user following another user, be it both individual user or corporate user (i.e. SonyMusic). A *relation* is to be formed when the users direct their messages at other users through a mention (syntax "@[username]") or retweet other users' tweets (syntax "RT @[username]"). Hashtag (syntax "#[word]") is used when users talk about a common topic.

For this project, the Twitter dataset given was dated from the time period 23 November 2011 to 26 February 2012, with the total of 65,533 tweets. Below is the simple summary of the characteristics of the dataset at hand:

User Levels

1. Individual: celebrities or any ordinary users who are involved
2. Corporate intermediaries: community-like users or corporate users, such as YouTube, Sonymusic, or a entertainment channel

Tweets

1. Directed: a user "mentions" another user/s in his/her tweet
2. Undirected: a user tweets without any mentions of another user

Relations

1. Interactive (mutual): users form a cyclic network by mentioning one other
2. Non-interactive (nonmutual): users may mention some other users who do not reciprocate

Definition of the Terms

Prior to the study, we have defined the following terms which are cricual and will be frequently used in our project:

Influence: the ability to transmit information to others or promote any form of activity, through one user's own behavior (RT, mention, hashtag, etc.) on Twitter.

Influencer

An influencer should have the following attributes[5]:

1. Reach: the ability to reach out to an audience. It can be measured through the user's popularity and proximity.
2. Resonance: the level of engagement with the audience. The frequency and period of the twitter contents may matter.
3. Relevance: having the relevant content, which is relevant to the topic. It is further defined as having the authority, trust, or affinity over a topic.

[5] Smitha, N. (April 2, 2014). How to define, identify, and engage social media influences for your brand. *Simply Measured.* Retrieved September 14, 2014 from http://simplymeasured.com/blog/2014/04/02/how-to-define-identify-and-engage-social-media-influencers-for-your-brand/

In a more quantifiable manner, an influencer is defined by the measure of centrality[6]:

1. Degree Centrality: Higher degree means higher number of contacts with other nodes
2. Betweenness Centrality: A node may occupy a "between" or intermediary position that connects many other nodes; it can be the center of information flow
3. Closeness Centrality: Higher closeness shows that the node is close to many others can quickly interact and communicate with them without going through many intermediaries.
4. Eigenvector Centrality: Eigenvector centrality measures the influentiality of the first node in the chain of nodes that subsequently influence many other nodes.

Assumptions Made for the Analysis

The data cleaning process has taken place over a couple of weeks, which was more than our expected duration, due to the complicated factors involved in Twitter data. Different assumptions have been made depending on which perspective we were looking at, and this resulted in different outcomes. Thus, it was vetted through carefully over time, until we came up with the final decision. As can be seen in the **Appendix 1**, there have been some conflicting issues due to the differing perspectives of the stakeholders involved.

Choosing the Population for Study

First question was to answer which population we wish to study in our analysis. Is it from the corporate or individual level? For the purpose of our project, we restricted our attention solely to individuals, in order to see the dynamics of how each individual can exert influence on other Twitter members. It is more accurate to see the level of influence of an individual on another individual (one-to-one), rather than a corporate on an individual (many-to-one). Also, for the future progress, it is easier to see the profiling of individual users rather than community users. For this, we will be narrowing down our target population to individual levels, not corporate levels.

Definition of Interactivity

It is clear that for our social network analysis, we are only using the directed tweets, to see the relation between one user to another. Directed tweets include the mentions (syntax @[username]), which shows that a user is "interacting" with another user.

Another issue arose with the definition of interactivity. The question was to either 1) include all the nodes even though they do not make cyclic relations or 2) include *only* the nodes that form cyclic relations and remove the nodes that end the cycle. This question on interactivity in Twitter users has been addressed, thanks to Rosenman[7] in his thesis paper "Retweets—but Not Just Retweets: Quantifying and Predicting Influence on Twitter."

---

6 Wasserman and Faust. (1994). Centrality & Prestige. Retrieved September 25 from https://www.soc.umn.edu/~knoke/pages/Centrality_and_Prestige.doc

7 Rosenman, E. (2012). *Retweets—but Not Just Retweets: Quantifying and Predicting Influence on Twitter*. Cambridge MA: Harvard College. Retrieved October 3, 2014 from http://www.eecs.harvard.edu/econcs/pubs/Rosenman_thesis.pdf

In his paper, he states that Twitter has an asymmetric following model, wherein users in Twitter form directed but not necessarily reciprocal relationships. This means that even though a user A follows a user B, without user B having to follow user A. Thus, these non-mutual relationships should be considered into our data cleaning, due to the nature of Twitter data. Thus, we have made our decision to include any nodes that make any kind—mutual or non-mutual—of relation.

Nature of Retweets (RT)

In a lot of Twitter analysis, RTs are removed due to the ambiguity it gives. However, we are including RTs in our analysis. Retweets are useful in determining influence, because a retweet requires another individual to read someone's tweet, be interested in the tweet content, and decide to share it with his or her own followers.[7] Thus, high number of retweets is definitely an indication of a user's influentiality.

The dataset contains different representations of Retweets, including:

Types	Tweet_from	Retweet_from	Content
RT	UserA	UserX	**RT** @UserX content content
Quotation marks	UserA	UserX	"@UserX: content content."
Via	UserA	UserX	This is so funny! **Via** @UserX content content content

Among the different representations, we have decided to use the symbol RT as the sole representation of Retweets in our analysis. The rest will be considered as ordinary mentions (syntax @[username]).

RTs are unique in different ways. Even though RT has a mention syntax, the relation is different from other ordinary mentions because the influence is inversed. For example, assume that UserA tweets:

UserA: Check this out!

The following scenario from another user UserB can happen:

1. UserB: @UserC @UserD have you heard of Kolaveri Di?
2. UserB: **RT**@UserA Check this out!
3. UserB: @UserC @UserD **RT**@UserA Check this out!
4. UserB: **RT**@UserA Check this out! @UserC @UserD

Scenario	Author_from	Author_to
1	B	C
	B	D
2	A	B
3	B	C
	B	D
	A	B
4	A	B

In scenario 1, it is quite straightforward that UserB is influencing UserC and UserD. In scenario 2, since UserB retweeted UserA's tweet, UserA was able to influence UserB to take the action to retweet. Scenario 3 is combination of Scenario 1 and 2. Scenario 4 is an exception because it is very ambiguous to determine if 1) UserB is mentioning UserC and User D or 2) UserC and UserD are part of the retweet. Thus, in cases of scenario 4, we have decided to include the immediate user tagged after RT, and remove the users who are tagged after.

Methodology

After finalizing the necessary assumptions, we cleaned the dataset accordingly and removed the corporate intermediaries. The usernames from directed tweets were extracted, for both author_from and author_to. Duplicates were removed to make a clean tabulated list of author_from and author_to.

Using the clean list of author_from and author_to, we ran NodeXL to plot a networked graph as well as to find out different metrics to measure the influentiality.

NodeXL Outcomes

1. Top 10 influencers for Betweenness Centrality

We understand that the higher the betweenness centrality score, means that the higher frequency where in a node flows between other notes on the shortest paths. They can be the intermediaries in transmitting information to other nodes and in good position to influence.

We found the top 10 users with the highest betweenness centrality as follows:

Username	Unstandardized Betweenness Centrality
dhanushkraja	19158259.578
anirudhravichan	7242922.889
cynthiacollins2	6334624.276
padhuu	4495465.369
ArjunArtist	3802142.363
linhgromleydtco	3461330.751
dallassharkoafy	2695680.535
ash_r_danush	2073457.574
precioushewlint	2040969.615
acthemc	1930521.657

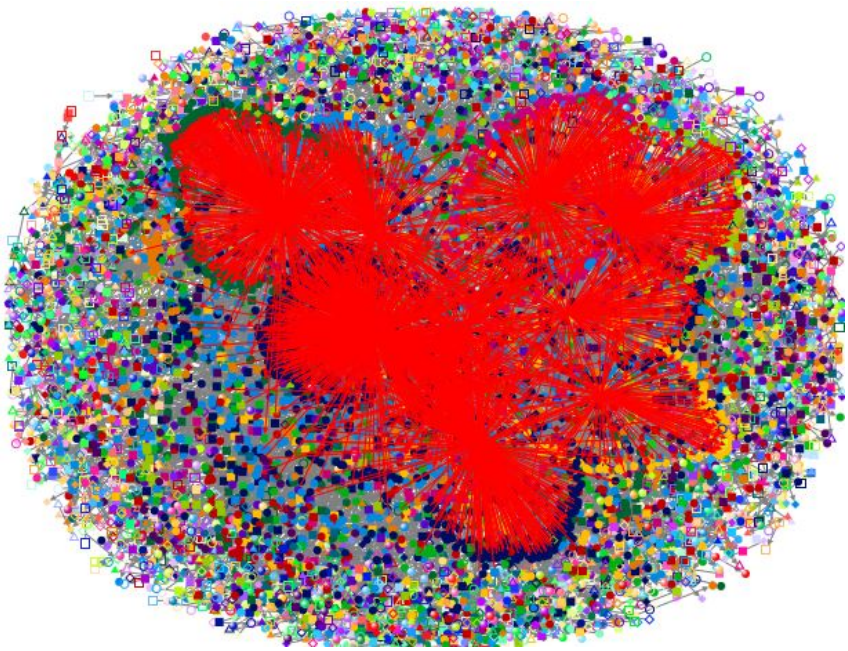


Figure 1. Top10 Betweenness Centrality

2. Top 10 influencers for Eigenvector centrality

Eigenvector centrality measures the influentiality of the first node in the chain of nodes that subsequently influence many other nodes. It is under the assumption that one node can affect other nodes simultaneously.

We found the top 10 users with the highest eigenvector centrality as follows:

Username	Eigenvector centrality
dhanushkraja	0.026
anirudhravichan	0.011
ash_r_danush	0.007
shrutihaasan	0.004
cynthiacollins2	0.003
padhuu	0.002
an1rudh99	0.002
nakkeeranteam	0.002
sri50	0.002
keerthinik	0.002

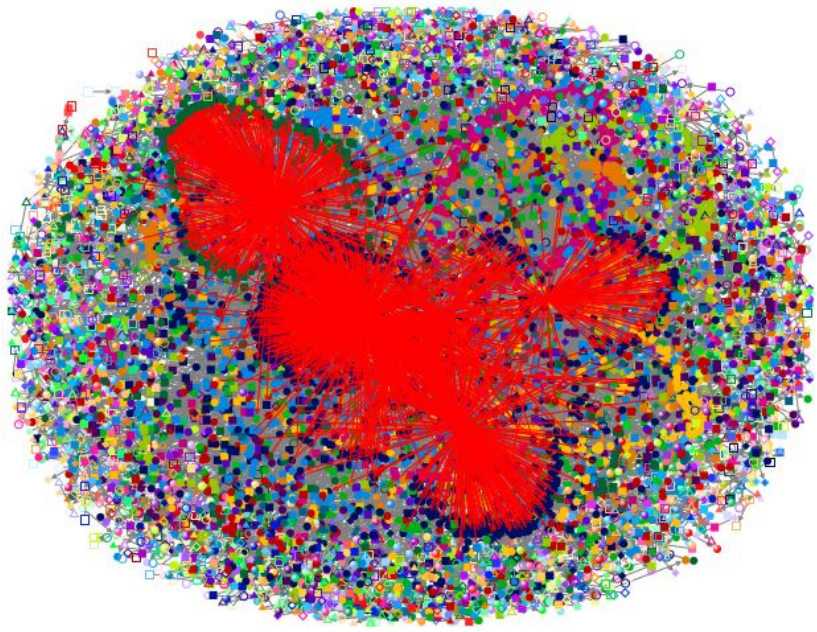
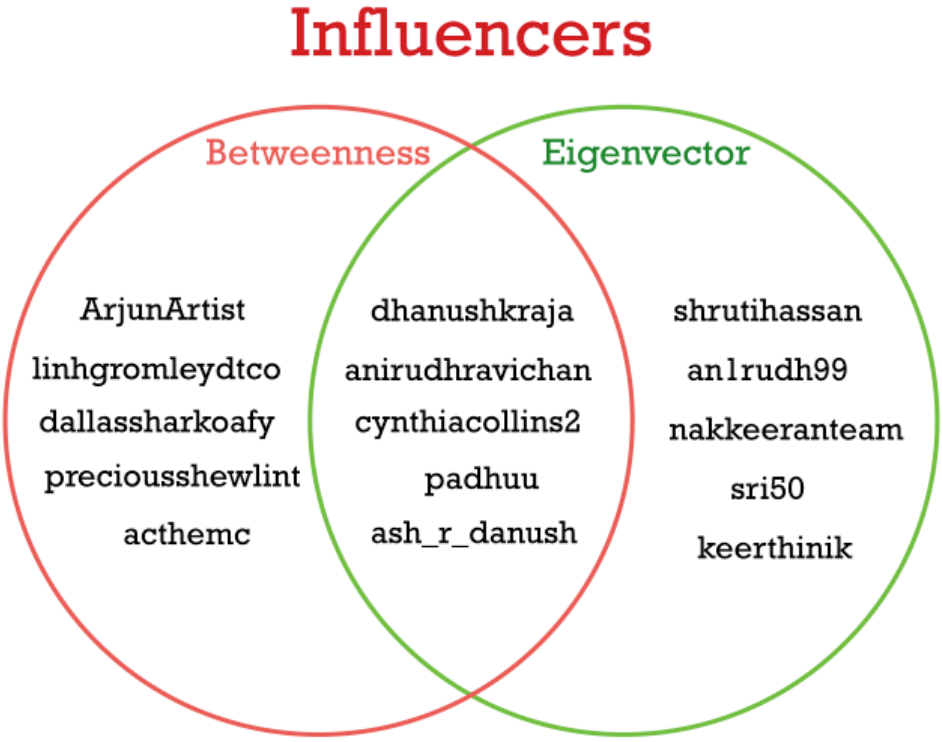


Figure 2. Top10 Eigenvector Centrality



Based on what we found on both centrality measures, here is a diagram to illustrate the influencers.

Influencers within the red circle act as intermediaries to transmit information to many other users, hence are the center of information dispersion.

Influencers within the green circle are the front runners in a chain of network. These influencers will start to transmit information, influencing their followers to follow subsequently, creating a chain of network.

The five influencers in the center are high in both betweenness and eigenvector centrality, making them the top influencers in the Kolaveri Di dataset.

Future Plans

Timeline

Week	Objectives and Tasks	Remark
7 (Midterm)	❖ Midterm presentation with Prof. Srini ❖ Adjust the direction of the project based on the feedback ❖ Prepare for presentation and report	
8	❖ Discussed subsequent plans for project ❖ Re-do data cleaning ❖ Sieve out keywords for intermediaries	
9	❖ Prepare for presentation and report ❖ Gather more insights on the top 10 users ❖ Start working with R ❖ Prepare data for time clustering	15th Oct - Midterm Presentation
10	❖ Finalize R ❖ Start working on time clustering	
11	❖ Come up with feasible marketing recommendations and conclusion of the project ❖ Finalize data on time clustering	
12	❖ Prepare for presentation and the completion of report	
13	❖ Final presentation ❖ Revise final report based on feedback	
14	❖ Submission of the final report	

Appendix 1: Timeline of Assumptions

Date	17 September 2014
Assumptions made	1. Include all the users (both individual and corporate level) as long as they mentioned someone OR are mentioned by someone else. 2. Only the tweets with mentions (@[syntax]) will be included in the social network analysis, because we're looking at who are interacting with who. 3. Only the RTs that are followed by a mention (@[syntax]) are included in the dataset for analysis. 4. Lists of "author_from" (the ones who create the tweet content) and "author_to" (the ones who are mentioned by author_from) are created, and duplicates are removed (if they have the same relation for more than one time).
Remarks	1. The graph generated from NodeXL is too huge, so it is very hard to detect the relationships between the actors. 2. Re-consideration on the inclusion of certain users.

Date	24 September 2014
Assumptions made	1. The node YouTube is removed from the dataset because YouTube does not make any interaction with other nodes. a. YouTube is only mentioned when users watch the video via YouTube, thus does not have any direct relation with other users. 2. Other corporate intermediaries (i.e. mtvindia, sonymusic_south) are included due to their direct relations with other nodes. a. Corporate intermediaries make direct relations by mentioning specific people or replying to others' tweets mentioned to them. 3. Only the nodes who "interact" or form cycles with other nodes are included. If one nodes ends the relations, they're removed because no network is formed from then on. a. We assumed that only the users who form reciprocal relations matter.
Remarks	1. NodeXL generated shows that the number of nodes and edges have been reduced. 2. The definition of "interactivity" or reciprocation was questioned.

Date	25 September 2014
	We met up with Yazhe who shared the possibilities of not removing any nodes (be in corporate intermediaries or the nodes that end the cycle) because some important information may be lost.
Remarks	1. The inclusion of corporate intermediaries is still not very confirmed.

Date	2 October 2014
Assumptions made	1. Nature of RT questioned: If person A re-tweet's person B's tweet, person B should have more influence on person A instead. This relation is inversed, as compared to other relations with pure mentions. a. For example, if Abby mentions Bob, (**Abby: @Bob**, check this out!) the relation becomes A influencing B. b. However, for RTs, if Abby retweets Bob's tweet (**Abby: RT "@Bob** Kolaveri Di is great"), the relation becomes Bob influencing Abby, not Abby influencing Bob. c. Assuming that there is a different set of influencers throughout different timings, time clustering may be important.
Remarks	1. We should re-think through the definition of influence, and adjust the RT relations accordingly. 2. We can consider time clustering towards the end of the project.

Date	9 October 2014
Assumptions made	1. Include ONLY the individual users, as we are planning to filter out who are the external influencers who made the video go viral. 2. In a lot of Twitter analysis, RTs are removed due to the ambiguity it gives. However, we are including RTs in our analysis. a. Retweets are useful in determining influence, because a retweet requires another individual to read someone's tweet, be interested in the tweet content, and decide to share it with his or her own followers (Rosenman, 2012) b. High number of retweets can be the indication of a user's influentiality. 3. RTs are included in our project, and are considered an inverse relation because the retweeted user is influencing the retweeting user due to its nature of influence. 4. As mentioned, only the tweets with mentions (@[syntax]) will be included in the social network analysis. 5. Lists of "author_from" (the ones who create the tweet content) and "author_to" (the ones who are mentioned by author_from) are created, and duplicates are removed. 6. We are including all users with non-mutual relationships, due to the nature of Twitter. a. In fact, only 22% of the relations in Twitter are mutual! (Rosenman, 2012)