On Modeling Virality of Twitter Content*

Tuan-Anh Hoang, Ee-Peng Lim, Palakorn Achananuparp,
Jing Jiang, and Feida Zhu

School of Information Systems, Singapore Management University

Abstract. Twitter is a popular microblogging site where users can easily use mobile phones or desktop machines to generate short messages to be shared with others in realtime. Twitter has seen heavy usage in many recent international events including Japan earthquake, Iran election, etc. In such events, many tweets may become viral for different reasons. In this paper, we study the virality of socio-political tweet content in the Singapore's 2011 general election (GE2011). We collected tweet data generated by about 20K Singapore users from 1 April 2011 till 12 May 2011, and the follow relationships among them. We introduce several quantitative indices for measuring the virality of tweets that are retweeted. Using these indices, we identify the most viral messages in GE2011 as well as the users behind them.

1 Introduction

1.1 Objectives

Events that attract social media attention often see their information becoming highly popular through word-of-mouth. These highly popular information are *viral*. Examples of viral information include viral videos, viral blog posts, viral games, etc.. Viral information may have accounted for most of the internet traffic consumption as they get pass on from people to people. Businesses are also trying to tap on viral information to promote brand and product awareness and this has resulted in several viral marketing techniques.

In this paper, we focus on viral content in Twitter. Twitter has seen very heavy usage in many countries due to its very low barrier to messaging posting and realtime dissemination. In Twitter, users *follow* other users to receive short text messages also known as *tweets* from the latter users. Tweets can be posted via mobile phone or desktop machine. To study viral content in Twitter during the Singapore's GE2011 event, we developed the *Palanteer* search engine[1] that offers tracking and search of GE2011 tweets. Palanteer performs daily crawl of tweets from 1 April 2011 till 12 May 2011 generated by users located in Singapore. The election date of GE2011 was May 7, 2011. Since GE2011 is a socially interesting

* We would like to acknowledge that this research was carried out at the Living Analytics Research Centre (LARC), sponsored by Singapore National Research Foundation and Interactive & Digital Media Programme Office, Media Development Authority.

[1] http://palanteer.sis.smu.edu.sg

C. Xing, F. Crestani, and A. Rauber (Eds.): ICADL 2011, LNCS 7008, pp. 212–221, 2011.
© Springer-Verlag Berlin Heidelberg 2011

event, we expect viral information to exists in the Twitter data generated by the event.

In this paper, we thus propose the following research objectives related to viral content and highlight the corresponding contributions:

- **Measurement of viral content:** One may measure viral information by their popularity, i.e., the number of users adopting the piece of information. Popularity however does not consider the word-of-mouth, or user-to-user propagation. It is therefore important to consider other quantitative measures of viral content. One major contribution of this paper is to introduce several new measures that use retweeting structure to quantify the virality of tweet information.
- **Analysis of viral content:** Using our proposed virality measures, we analyze the GE2011 Twitter data. We looks out for viral tweets propagated among Singapore based users, as well as viral politicians and topics found in these tweets. This allows us to understand the foci of attention of users in the event.
- **Measurement and analysis of viral behavior:** Viral information can be associated with users who generate them and users who propagate them. Users who generate viral information are likely influential in the Twitter network. In this analysis, we focus on measuring and analyzing these users.

There are several useful applications of virality measures and analysis. Viral marketing techniques can apply virality to measuring the success of viral advertisements or new product diffusion within the Twitter network. By tuning the advertisements and changing product designs, one may improve virality which in turns improves the marketing outcome. Search engines can utilize virality to rank information so as to return more socially interesting search results. Finally, we may use virality to identify influential users in the Twitter network related to some social events. These may also be the users one should track closely for future events.

2 Related Works

The term virality originated from medical science to describe the ability of viruses to spread among organisms. Emerged from economics and social studies, viral behavior of an object is one that indicates how good it is at "network-enhanced word of mouth"[5]. Marketing community later created viral marketing strategies that focus on designing messages that grow awareness of products or advertisements within a targeted customer community[4,11]. These previous works focus on largely qualitative study instead of quantitative modeling. The effects of network is also often neglected.

Leskovec, Adamic and Huberman analysed the viral effect of customers recommending products in a large user-user recommendation network over time[6]. They observed that most customers do not propagate product recommendation

after purchases. Only very few products enjoy active recommendation within small communities. The recommendation network is quite different from Twitter network which involves little cost for users to generate and propagate tweets. The work also did not introducing any quantitative measure. Ienco, Bonchi and Castillo proposed to rank messages for users to consume so as maximize the overall user activeness in the network[4]. They however focused on the optimization of the selection of a small subset of messages instead of measuring the message virality.

3 Twitter Dataset

Data crawling. Using the *Palanteer* search engine, we collected around three weeks of Twitter data before the Singapore's general election on May 7, 2011. The data crawling selected a seed user set U_0 which contains 58 Twitter users who have dominant interests in GE2011. They include the political party accounts (e.g., `PAPSingapore` and `wpsg`), politician accounts (e.g., `georgeyeo`) and political commentators (e.g, `temasekreview`). We then derived the followers and followees of users from U_0 and created a larger set of users denoted by U. All tweets generated by U during the three weeks are crawled.

Dataset selection. As there are tweets in the crawled data that are not related to GE2011, we hence extracted a set of tweets generated by U containing the 5 most frequently mentioned politician names during the election, namely, George Yeo, Tin Pei Ling, Nicole Seah, Chiam See Tong and Low Kia Thiang. The first two are members of the ruling party PAP while the latter three belong to opposition parties NSP, SPP and WP respectively. The statistics of this dataset (known as the 5-Politician dataset) is shown in Table 1:

Table 1. Statistics of 5-Politician Dataset

# tweets	# users	# retweets	# retweeting users
32,696	4718	11,436	2994

Inferred Retweets. In the 5-Politician dataset, we realised that many retweets are in response to tweets not covered in the dataset. For example, our dataset contains a tweet with content, ``RT @djkoflow: NSP at lavender foodcourt!!! Nice will i see nicole seah?'' which is a retweet of another tweet by `@djkoflow` not found our database. In this case, we inferred that the tweet ``NSP at lavender foodcourt!!! Nice will i see nicole seah?'' actually exists based on the former tweet. There is also an inferred retweeting relationship between these two tweets. By introducing inferred retweets, we increased the number of retweets by 48% from 11,436 to 16,932. We define an *original* tweet to be one that is not a retweet. We obtained 4096 original tweets that have been retweeted. The increase in retweets allows us to have a more complete set of data for modeling the virality of tweet messages.

4 Virality Models

In this section, we present some models for measuring the virality of tweets, topics, and users. Each model is derived based on a different set of heuristic principles. Each model assigns a numerical virality value to an entity of some type. We use $mscore(m)$, $tscore(t)$, and $uscore(u)$ to denote the virality value of the tweet m, topic t, and user u respectively. A standard set of notations used in the definition is given in Table 2.

Table 2. Notations

$M/T/U$	set of tweets/topics/users
$u(m)$	original author of the tweet m
$U_R(m)$	set of users retweeting the tweet m
$W(m)$	set of non-stop words in tweet m
$M(t)$	set of (original) tweets in topic t
$M(u)$	set of (original) tweets tweeted by user u
$M_R(u)$	set of (original) tweets retweeted by user u
$Follow(u)$	set of followers of user u
$mscore(m)$	viral score of tweet m
$tsore(t)$	viral score of topic t
$usore(u)$	viral score of user u

4.1 Viral Tweets

Virality by retweet count. Like viruses, a viral tweet should be one having ability of being replicated on as many people as possible. In the Twitter context, retweet is analogous to replication. Therefore, it is natural to use the frequency of a tweet being retweeted to measure its virality. Since Twitter allows a user to retweet a tweet at most once, the number of times a tweet is retweet is simply the number of users retweeting it. That is,

$$mscore_{rtc}(m) = |U_R(m)| \tag{1}$$

Virality by retweet likelihood. The retweet count model tells us the popularity of a tweet, but does not tell us how good the tweet is propagated through the social network structure. Consider the two original tweets, m_1 by user u_1 and m_2 by user u_2. Suppose that only 4 out of 20 followers of u_1 retweeted m_1, and all 3 followers of u_2 retweeted m_2. By retweet count model, m_1 is more viral than m_2. However, one may consider m_2 to be more viral since all of its receivers have retweeted while only a small fraction of users receiving m_1 actually retweet m_1.

We therefore propose an alternative model that measures the virality of a tweet m by its retweet likelihood which is computed by the fraction of users retweeting m over all users receiving m.

$$mscore_{rtl}(m) = \frac{|U_R(m)|}{|\bigcup_{u \in U_R(m) \cup \{u(m)\}} Follow(u)|} \tag{2}$$

Note that our dataset was crawled from a subset of Twitter users at most one hop away from the set of 58 core users in the follow graph. The dataset hence does not have complete follower information for all the non-core users. Given this incomplete follow information and that follow links are rather weak relationships, we decide to use the set of users interested in u instead of the set of u's followers. Furthermore, a user can may receive a tweet by u through public timeline and generate a retweet even when he/she does not follow u.

We say that user u_1 is interested in user u_2 if u_1 mentions "@u_2" in at least k of u_1's tweets. In our experiments, we set $k = 1$. This means u_1 mentions u_2 at least once or u_1 retweets at least one tweet from u_2.

The Equation 2 is now modified as follows:

$$mscore_{rtl}(m) = \frac{|U_R(m)|}{|\bigcup_{u \in U_R(m) \cup \{u(m)\}} \{u' : u' \text{ interested in } u\}|} \tag{3}$$

Virality by a combined model. The virality by retweet likelihood model has a shortcoming since it favors tweets published by users having fewer followers while ignoring the tweet popularity among retweeting users. Therefore, we propose a combined model incorporating both retweet count and retweet likelihood.

$$mscore_c(m) = mscore_{rtc}(m) \cdot mscore_{rtl}(m) \tag{4}$$

4.2 Viral Topics

Topic construction. A topic is a group of tweets of similar content. We assume that each tweet contains only one topic and adopt a clustering approach to construct topics. We take a modularity based clustering algorithm [3] on a *tweet graph* so as to derive subgraphs representing topics. The tweet graph consists of original tweets as nodes, and pairs of tweets with overlapping terms as edges. Each edge is weighted by the similarity of the two corresponding tweets. To compute similarity between two tweets, we use the "bag-of-words" representation of each tweet after removing tweet-encoding terms, e.g. *@*, *RT*, and *via*, common stop words and internet slang words[7]. The remaining words form a word vocabulary W. The similarity between tweet m_1 and tweet m_2 is computed by Jaccard coefficient as follows. At the end of clustering, we obtained 69 topics.

$$similarity(m_1, m_2) = \frac{|W(m_1) \cap W(m_2)|}{|W(m_1) \cup W(m_2)|} \tag{5}$$

Note that $similarity(m_1, m_2)$ returns a value between 0 and 1. We remove from the tweet graph all edges with weight less than 0.25 to reduce noise in topic construction.

Modularity $q(C)$, proposed by Newman[9], is an objective function that measures the quality of cluster results C of a network. $q(C)$ falls within the range $[-1, 1]$, and the larger $q(C)$ the better is the resultant clusters C. Maximization of $q(C)$ is proven to be a $NP - hard$[1] problem. Some top-down and bottom-up heuristic and approximate algorithms for this problem were proposed by

Newman et. al [9], [3], [8]. In our work, we choose the agglomerative method to determine clusters as topics since it outperforms the top-down method as reported in [3].

Topic virality. We employed a simple principle to measure virality of each topic. Topic with a lot of viral tweets is viral. Then, the virality of topic t is simply the sum of virality of all tweets in t. That is

$$tscore(t) = \sum_{m \in M(t)} mscore(m) \tag{6}$$

Similar to modeling the virality of tweet, we derive three models to measure topic virality based on three different tweet virality models. Due to space constraint, we shall just show the results using the combined model, i.e., $mscore = mscore_c$.

4.3 Viral Users

There are two basic modeling approaches for viral users, namely the *author-take-all* and the *shared contribution* models.

"Author take all" model. This approach measures the virality of a user by how viral her tweets are. In this model, we consider the virality of a tweet can be fully attributed to its original author but none to the users retweeting the tweet. In other words, retweeting users does not get any share of the virality of the tweet. With this assumption, the virality of user u is the sum of virality of all original tweets of u. That is,

$$uscore(u) = \sum_{m \in M(u)} mscore(m) \tag{7}$$

Depending on how we measure tweet virality, we can have three corresponding ways to define the virality of user. In this paper, we shall only present the one using combined virality model.

$$uscore_c(u) = \sum_{m \in M(u)} mscore_c(m) \tag{8}$$

"Shared contribution" model. Sometimes, a tweet may not be viral until it is retweeted by an influential user who attracts a large number of subsequent retweets by her followers. To account for the contribution of retweeting users to tweet virality, we propose the *Share contribution* model that measures the virality of a user u by including u's contribution to all tweets' virality (including retweets). Given an original tweet m and a user u, the contribution of u to virality of m is proportional to number of retweets on the tweet (if u is author of m) or on the u's retweet of m (if u retweets m). That is,

$$uscore(u) = \sum_{m \in M(u) \cup M_R(u)} \left(\frac{mscore(m)}{1 + |U_R(m)|} \cdot (1 + |U_R(m) \cap Follow(u)|)\right) \tag{9}$$

Similarly, we have three ways to compute the virality of users corresponding to three models mearsuring tweet virality. Furthermore, as mentioned in Section 4.1, we slightly modify Equation 9 by using the users interested in u instead of u's followers. In the following, we show the one using the combined model.

$$uscore_c(u) = \sum_{m \in M(u) \cup M_R(u)} \frac{mscore_c(m)}{1 + |U_R(m)|} \cdot (1 + |U_R(m) \cap \{u' : u' \text{ interested in } u\}|) \quad (10)$$

5 Data Analysis

We now apply our virality measures on the 5-Politican dataset. Our aims is two-fold. Firstly, we want to analyze the differences between measures for tweet messages, topics and users. This comparison helps us to understand the characteristics of these measures. Secondly, we want to determine the highly viral messages, topics and users in GE2011 and provide some empirical analysis on them. The analysis will lead us to understand the socio-political issues that interest the Singapore electorates. In some cases, these issues may not appear within the radar of traditional mainstream media.

5.1 Tweet Analysis

According to the combined virality measure $mscore_c$, the following are the most viral tweets that have been both retweeted by many users and have experienced high retweet likelihood. The combined virality measure is adopted due to its balanced consideration of both retweet count and likelihood. The $mscore_c$, number of retweets each of them had generated and the retweet likelihood are shown in normal typeface, boldface and italized numbers respectively before each tweet.

1. (7.50, **127**, *0.67*): `Retweet if you want George Yeo to remain in Parliment. #sgelection`
2. (7.41, **206**, *0.41*): `Join us in our salute to George Yeo and Chiam See Tong for their service to Singapore. RT this. #sgelections #sgelection`
3. (6.46, **172**, *0.43*): `Retweet if you are sad that Tin Pei Ling is elected more than you are sad that Nicole Seah is not elected. #sgelections`
4. (5.98, **161**, *0.42*): `Tin Pei Ling is in Parliament..God Bless Singapore`
5. (4.58, **138**, *0.38*): `Retweet this if you love Mr Chiam See Tong, the people's humble leader who sat in his aluminum cubicle all these years for his people.`
6. (3.61, **48**, *0.86*):`If only we can exchange Chiam See Tong and George Yeo for TPL and WKS:( #sosingaporean`
7. (2.43, **69**, *0.40*): `I'm happy that Tin Pei Ling in our govt. Now everyones truly represented. Even the intellectually disabled :) #sgel ...`
8. (2.21, **85**, *0.30*): `To all those regretting the loss of George Yeo, Chiam ST, the election of TPL, remember: it's the GRC system that caused ...`
9. (1.86, **108**, *0.20*): `With extra $16000 per month in her pocket now Tin Pei Ling can upgrade from Kate Spade to Hermas. #GeorgeYeoInTinPeiLingOut #SGElections`
10. (1.68, **28**, *0.68*): `@Xiaxue my 4yr old daughter is gonna be so sad when she knows that George Yeo, the uncle that gave her a lollipop, lost.. ;(`

All the above viral tweets were generated right after election results were known and they expressed the sentiments from the users, including giving encouragement to George Yeo and Chiam See Tong, the two popular politicians who lost their parliamentary seats after election, and satirical remarks to Tin Pei Ling who suffered from online flaming throughout her election campaign. In particular, the topmost viral message shows support to George Yeo despite his loss.

Among the top 10 tweets, one can find some with not many retweets (28 retweets) but high retweet likelihood. This suggests that such tweets attracted attention of followers of every user who tweets or retweets them. For example, the 10th most viral tweet, ‘‘@Xiaxue my *(omitted)*’’, has observed a total of 41 users receiving it, out of which 28 have retweeted. This is a significantly high retweet likelihood making the tweet very viral.

On the other hand, we can also find another tweet ‘‘WP’s 5-man Aljunied team have been garlanded at Hougang Stadium. Low Thia Khiang thanking supporters #sgelections’’ having similar number of retweets (33 retweets) but much lower retweet likelihood. This resulted in the tweet being ranked 1217th, much lower than the earlier example. The low retweet likelihood is caused by a total of 806 users receiving the tweet but only 33 retweeted. In other words, users are generally not interested to pass on this retweet to their followers.

5.2 Topic Analysis

We next analyze the viral topics in 5-Politician dataset using the model proposed in Section 4. The combined message virality model is used in deriving the topic virality measure. We rank all the 69 constructed topics by decreasing order of virality values and show the top 10 viral topics in Table 3. The top viral topics are dominated by topics with large number of original tweets and retweets.

Even among the top viral topics, the top five appear to have significantly larger virality scores compared with the lower five. This observation is also supported by the number of original tweets and number of retweets. The topmost topic is largely about encouraging George Yeo to stand as a Presidential candidate upon the loss of his parliamentary seat. A total of 1115 original tweets followed by 4600 retweets were generated on this topic. The second most viral topic covers the suggestion to replacement of Tin Pei Ling who won the election by George Yeo. Again, this topic generates many original tweets and retweets. Similar observations can be made for the remaining viral topics.

5.3 User Behavior Analysis

Now, we analyse the virality behaviors of users in our dataset. Table 4 shows the top 10 viral users by virality scores using author-takes-all (ATA) and shared contribution (SC) models. The two models return different but overlapping sets of top users. The ATA model favors users who contribute many highly viral messages while the SC model favors users who also contribute to retweets of viral messages. The highly viral users include information aggregator users (e.g., `yahoosg`, `sgelection`), mainstream media (e.g., `stcom`, `todayonline`, etc.),

Table 3. Top 10 Viral Topics

	Virality Score	# Orig Tweets	# Re tweets	Topic Words
1	23.65	1115	4600	lost,yeo,wp,parliament,chiam,foreign,minister,george,tpl,president, pap,aljunied,grc,tong,sad
2	14.94	828	4077	ling,yeo,pap,lings,mp,goh,spade,george,say,pei,dont,kate,tong,tin,parliament
3	10.92	812	2905	ling,rally,nicole,party,pap,nsps,facebook,tpl,pei,nsp,tin,marine,grc,seah,parade
4	8.62	442	2009	tong,lost,chiam,mr,pasir,you,singaporeans,potong,residents,payoh,respect, crowd,spps,bishantoa,spp
5	2.01	253	1127	wps,thia,low,wp,lim,mao,rally,khiang,yeo,hougang,george,chen,sylvia,grc, aljunied
6	0.95	36	98	did,yeo,salute,mr,ukiss,you,proud,mp,minister,respect,george,election, grc,elections,aljunied
7	0.75	16	132	ling,truly,government,disabled,represented,happy,bwahaha,tt,pei,tin,govt,every, im,everyones,intellectually
8	0.41	29	101	selfblinged,statement,hes,speech,thing,poor,minister,press,yeos,conference,mr, george,commentary,yeo's,why
9	0.29	7	90	really,yeo,wp,politics,hua,lim,quit,announced,george,me,consider,hwee,needs, important,decision
10	0.28	2	89	bts,yeo,wong,getting,kss,still,mistakes,george,tpl,sad,mah,weight,parliament

Table 4. Top 10 Viral Users

User	Rank/Score (ATA)	Rank/Score (SC)
yahoosg	1/90.1	6/246.4
firefliesinajar	2/85.7	
thenooselite	3/80.1	
temasekreview	4/73.7	
fakemoe	5/70.0	2/535.9
merylzhanyee	6/54.5	
stcom	7/49.1	4/284.8
todayonline	8/44.6	1/768.8
sosingaporean	9/41.1	
fake_pmlee	10/39.1	
sgelection		3/419.1
tocsg		5/251.6
ongweizhong		7/237.2
eisen		8/184.3
mrbrown		9/179.5
calvin_lee		10/136.5

satirical users (e.g., `fakemoe` and `fake_pmlee`), and activist users (e.g., `mrbrown`, `thenooselite`, `ongweizhong`, `calvin lee`, etc.). The user `firefliesinajar` ceased to exist at the point this paper was written. Interestingly, the mainstream media users have demonstrated the ability to stay viral compared with other users by generating viral tweets.

6 Conclusion

In a social network, messages passed among users may become viral especially in an event that attracts wide interests. To offer a systematic approach to study virality of content, this paper proposes a few models to measure virality at the tweet, topic and user level. Other than considering the number of retweets of a tweet, we also introduce retweet likelihood as another component in these

models. Using the Palanteer search engine, we collected a sizeable collection of Twitter data and applied our proposed models on it. The empirical analysis results demonstrated that the virality models can effectively find the viral content and users. As part of future work, one can extend the virality models for searching Twitter content. This will allow viral content to be returned as search results to meet the users' demand for trending content in social media. Viral messages may be linked to influential users in the network although this relationship has yet to be well studied. Several researchers have studied different ways of calibrating influential users [10,2]. It is thus interesting to study the relationship between influential users and users contributing to viral content.

References

1. Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Trans. Knowledge and Data Engineering 20(2), 8577–8582 (2008)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: The million follower fallacy. In: ICWSM (2010)
3. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70 (2004)
4. Ienco, D., Bonchi, F., Castillo, C.: The meme ranking problem: Maximizing microblogging virality. In: ICDM Workshops (2010)
5. Jurvetson, S.: What exactly is viral marketing? Red Herring 78, 110–112 (2005)
6. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM TWeb 1(1) (2007)
7. Manning, C., Raghavan, P., Schatze, H.: Introduction to information retrieval (2008)
8. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74 (2006)
9. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA, 8577–8582 (2006)
10. Weng, J., Lim, E.-P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential twitterers. In: ACM WSDM (2010)
11. Wilson, R.F.: The six simple principles of viral marketing. Web Marketing Today (2005)