10-2011

Virality Modeling and Analysis

Tuan Anh HOANG
Singapore Management University, tahoang@smu.edu.sg

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

Virality Modeling and Analysis

Tuan-Anh Hoang, Ee-Peng Lim
School of Information Systems
Singapore Management University
Email: tahoang.2011@phdis.smu.edu.sg, eplim@smu.edu.sg

***Abstract*—Virality is a virus-like behavior that allows a piece of information to widely and quickly diffuse within the network of adopters through word of mouth. It is about how easy users propagate information to their friends and friends of friends by means of diffusion. While virality of information has several interesting applications, there are much research to be conducted on virality. These areas of research include understanding the mechanism of virality, modeling the virality both qualitatively and quantitatively, and applying virality to applications such as marketing, event detection, and others. In this paper, we survey existing works on quantitative models for virality and the relationship between virality and other behaviors. Through the survey, we hope to offer common framework to study topis related to information virality.**

I. Introduction

Originated from medical science, the term virality was first used to describe the ability of viruses to spread from one host to another within a community of people. Subsequently, virality attracts the attention from researchers in marketing science, social sicence, and economics for various reasons. Jurvetson and Draper introduced the term *viral marketing* to describe the strategy used by Hotmail[1] to market its free email service [13]. Later, Jurveston defined *viral marketing* as "network-enhanced word of mouth" [12]. Although there are some controversies in Jurvenson's definition found in marketing science literature, e.g. disagreements about differences between viral marketing and word of mouth in works of Pastore [20], Helm [10], Modzelewski [18], Kirby and Marsden [14], and Ferguson [8], virality is widely understood as a social phenomenon that is strictly related to interpersonal communications. In research of item adoption and information diffusion, virality refers the ability of items in widely and quickly diffusing in a community through word of mouth. It is also used to indicate the ability of people in making viral items as they adopt these items. Such people include celebrities and sport personalities.

With important applications in both business and social science, there has been a number of research projects on understanding the mechanisms of virality. Several early qualitative and macro models for virality were proposed in marketing science [23], [1]. With the recent popularity of Internet and social network sites, virality may for the first time be quantitatively measured over a large network of users. In this paper, we survey the recent works on virality modeling and analysis with a focus on the computational aspects. We focus on the multiple facets of virality of an information item and how the facets correlate with effects of user network and content of the item itself.

The rest of this paper is organized as follows. In Section II, we present a taxonomy of virality related measurements. Works on understanding the relationship between virality and network behaviors as well as content behaviors are introduced in Section III and Section IV respectively. We present the applications of virality in Section V. Finally, we conclude and introduce our future research directions in Section VI.

II. Virality modeling

The virality of an item which can be a piece of information or some product is about how widely and quickly the item diffuses. Therefore, *number of adopters*, also known as popularity of the item, and *rate of adoption* are traditionally used to measure virality. Jurveston used a power model to measure the number of users adopting Hotmail service based on rate and quality of communication between adopters and their friends [12]. Hoang *et. al.* used retweet count to indentify most viral tweets in a socio-political snapshot of Twitter[2] [11]. Virality based on the rate of adoption was first proposed by Leskovec *et. al.* which measures the difference between the time when a user adopts the item and the time when the user receives the last recommendation about the item [16]. Later, Janghyuk *et. al.* modeled this time lag using Cox regression with additional predictive variables including the personal information of users sending and receiving item recommendations [15]. On the other hand, Wu *et. al.* characterized the virality of a piece of news content by its decay rate [28].

In the context of Youtube, Broxton *et. al.* postulated that not all popular videos on Youtube are viral [4], as popularity and rate of adoption are not sufficient postulated that virality. Guerini *et. al.* showed that popularity is only one of multiple facets of virality [9]. The other facets are:

- Appreciation: How much people *like* the item, e.g. the number of times the *like* button in Youtube[3] is hit for a particular video.
- Spreading: The likelihood that users will share (propagate) the item once they adopted, e.g. the fraction of followers retweeting a tweet in after they receive the tweet from their followees [11], or the likelihood that a user will adopt the item once her friends adopted [24].

[1] www.hotmail.com

[2] www.twitter.com

[3] www.youtube.com

- Buzz: The likelihood that users will give comment about the item, e.g. the ratio between number of comments and number of view counts of each video on Youtube
- Raising discussion: How much users discuss about the item, e.g. the number of subsequent comments about the item, including comments replying another comment.
- Controversiality: Ability of the item in splitting users into groups of different opinions, e.g. the ratio between numbers of times the item gets *like* and *dislike*.

While the above models assign for each item a virality score which falls in a range, say 0 to 1, a number of works determine virality by a binary indicator, i.e. items are viral if their behavior(s) exceeds a certain threshold, e.g. a video is viral if it has at least 10 millions view counts [26] or it is in the top most viewed videos [4].

Apart from works on items' virality, others go further to examine the virality of individuals in a user network. While virality of an item models how easy the item diffuses through the network, the virality of a user has to consider how easy the user propagates the item to her friends and friends of friends once she adopted the item. In [11], we proposed measuring the virality of Twitter users by their cumulative contribution to the virality of tweets they tweeted or retweeted. Contribution of a user to a particular tweet is computed by virality of the tweet if the user is the original author of the tweet, i.e. user who first tweeted the message, or by virality of the tweet weighted by the fraction of retweets due to retweet of the user. Janghyuk *et. al.* studied user virality by conducting a quasi-experiment where participants may send recommendation about an item to their friends. They proposed to compute virality of a user by the number of unique friends the user sent recommendation to after adopting the item [15].

III. Virality and network effects

Behaviors of the underlying user network were shown to have effects on the diffusion of innovations through the population of users [5]. In this section, we discuss some recent empirical researches on understanding how the different perspectives of virality correlate with behaviors of the user network.

The relationship between the popularity and the subnetwork of the first adopters were studied by Romero *et. al.* [24]. They found that these subnetworks of hashtags in more popular category, e.g. political hashtags, are more dense than others. Broxton *et. al.* characterized the subnetwork between all adopters of a Youtube video by the video's socialness [4]. The socialness of a video is determined by the fraction of social views, i.e. views from other websites where the video is embedded or by directly putting its url to web browser, in the first 30 days after the videos is posted. They found that there are differences in viewership patterns of highly social videos and less social videos, e.g. the former is more rapidly rise to and fall from the peak than the latter, the former has a bigger increasing in the rate of sharing at the time of the peak than the latter. On the otherhand, they also reported that not all social videos, i.e. videos have fraction of social views at least 80%, are in top 1% of most popular, and in the opposite way, only 21% of most popular videos are social. This means the popularity of a video may not due to the social effects.

In Twitter, effects of the subnetwork around the original author on the likelihood that her tweets are retweeted was studied by Petrovi *et. al.* [21]. Using a online learning model on the stream of tweets, they found that the number of times a user is listed[4], the numbers of the user's followers and friends are the most positively correlated features, while the number of time the users was mentioned and the number of tweets the user posted are the most negatively correlated features.

The dynamic of effects of the whole network on the virality was studied in the work of Szabo and Huberman [27]. They made a statistic on the fraction of influenced adopters, i.e users adopting a the item after their friend(s) adopted. This statistic shows that when the item raise to a certain degree of popularity, the fraction of influenced adopters significantly decreases. This means the underlying user network does not have much effect on how the item diffuses once it has obtained a certain degree of popularity.

Janghyuk *et. al.* studied how user's virality varies over a large network of recommendation in a viral marketing strategy [15]. They characterized user's virality by *speed* (the time that user's friends need to respond recommendations received from the user) and *volume* (the number of recommendations that the user sent to friends) which are respectively modeled by Cox regression [6] and negative binomial distribution process based on a set of predictive variables including the user's personal information, e.g. gender, age, and geology distance between the user and the receivers, and the number of recommendations was sent by the user or received. From learning the parameters for those models, they found that the differences between the sender and the receivers have a strong correlation with speed of virality of the sender.

IV. Virality and content

Research has shown that there is a relationship between virality and the content of messages. We expect the more interesting the information item is, the more users adopt the item. In this section, we present works on the rrelationship between content of information items and their virality.

Berger *et. al.* showed the strong effects of content on virality of online contents [2]. Hansen *et. al.* studied how the sentiment of a Twitter message affects the probability that the message is retweeted. They later classified tweets into news and social messages using a Naive Bayes classifier. They also computed *valence* and *arousal* scores of every tweet based on sentimal scores of non-stop words contained in the tweet. The likelihood that a tweet is retweeted is then modeled as a generalized linear function of *valence* and *arousal*. By learning parameters in the model, they found that the negative sentiment does not promote retweeting in random samples of all the tweets, while the negative sentiment does promote retweeting of news tweets.

[4]Twitter allows users to organize friends into groups according to some criteria, e.g by topics or by relationships. Each of such group is a *List*

Based on a social theory that a content is either viral or not, and its virality does not significantly depend on how much its adopters influence their friends, Guerini *et.al.* claimed that the virality of a content is strictly connected to the nature of the content [9]. They conducted experiments on a dataset collected from Digg [5] to show how the facets of virality may be predicted from content. They represented content of an item by its set of words which are tagged using PoS labels [6] and, for each facet, the item is considered viral or not if its measurement by means of the facet exceed or fall belew certain thresholds. They reported that, using SVM-light [7] with default settings, facets of virality can be effectively and independently be predicted. In the same approach, the work of Szabo and Huberman implicated that the virality of a Youtube video does not change over time [27]. They used an empirical study to show the strong correlation between the long term popularity of items (measured by number of view counts in Youtube or number of votes in Digg in the first 30 days after items are posted) with their early patterns of access (measured by number of view counts in first hour and in 7 days after posting in Youtube and Digg respectively).

To understand the temporal patterns of virality, Romero *et. al.* studied how the most widely used Twitter hashtags spread over time [24]. Using a large dataset of tweets, they extracted the top 500 most used hashtags and then manually classified them into 8 classes including *Celibrity*, *Games*, *Idioms*, *Movies/TV*, *Music*, *Political*, *Sports*, and *Technology*. Based on the user network inferred from *mention* relationships between tweets and users, they characterized the dynamics of the likelihood that a user adopts a hashtag by the curve showing how the likelihood varies as the number of the user's friends using the hashtag increases. They then computed *stickiness* (the value of the curve at the point that the curve attains maximum) and *persistence* (ratio between area of the field under the curve and area of the retangle with length is maximum number of influencing user and height is stickiness) of each hashtag, and compared mean and variance of persistence and stickiness of hashtags in different categories. They found that hashtags of controversial topics are more persistent. Wu *et. al.* studied the relationship between the decay rate of a piece of news with its content using a set of webpages whose *bit.ly* shorten URLs[8] were mentioned in Twitter [28]. They first indentified two classes of news based on their *decay time*, i.e the number of hours after the peak when the number of mentions first reaches 75% of the total. One class consists of persistent news whose decay time at least 24 hours, as suggested in [17], and the other consists of news with decay time less than 6 hours. They then analyzed emotion of content in each class using Linguistic Inquiry and Word Count (LIWC) [9] and extracted the trending words of each class. They found that news containing words with positive emotion and words related to art, advertising, and online marketing are more persistent than those with negative emotion and media related words; and news containing more words related to actions and tense are more rapidly decay.

Besides the content of an information item, the way adopters consumed the item also has a strong correlation with its virality. Shamma *et. al.* studied the relationship between the popularity of a video (measured by view counts) and the way it is shared through a synchronous environment [26]. Based on the digital traces collected from Yahoo! Zync, a plug-in for Yahoo Messenger that allow users to view and interact with a video simultaneously during chat sessions, they tried to predict whether a video has more than 10 millions views using a Naive Bayes classifier. Digital traces of a video in a chat session where the video is shared includes session related features, e.g. the users, the number of chat lines and words during the time that the video is played, and event related features, e.g. the number of start, stop, play, pause, fast forward, and rewind commands. The feature vector representing a video is aggregated from all the digital traces of the video.

On the user side, the virality of a user was shown to have strong correlations with the user's personal information [15]. Additionally, Rowe used a multiple linear regression model to find the relationship between Youtube users' popularity and other effects [25]. In this work, the popularity of a user is measured by the number of subscribers and the set of predictive variables includes both network related features, e.g, the number of subscriptions, and content related features. e.g. the numbers of post counts and user view counts, and the number of favourite counts. Their experiments running on 200 videos randomly selected from 2000 most recently uploaded ones (at the time of data crawling) show that the number of views of videos uploaded by a user and the number of favourite counts she created are most strongly correlated with her popularity.

V. Application

Based on the relationship between virality and other behaviors, one may use the virality in prediction and anomaly detection tasks. Ratkiewicz *et. al.* studied the problem of detecting political astroturfs (progation of memes containing untrue information by a single user or organization but disguised as the reaction of independent communities to some political entity, e.g. a politician, political group, product, service or event) based on the way they were propagated (this is different from spam detection where we focus on content of the message [22]. Based on a set of manually labeled memes they trained binary classifiers using AdaBoost and SVM models. The set of 31 features to represent a meme includes both network features (the statistics on diffusion network of the meme) and sentiment features (e.g. mood scores of the meme computed as in [3]). They found that the network related features are the most discriminative features. This suggests that the virality of a meme, which is here characterized by the subnetwork between adopters, could be used as a reliable indicator for quality of the content. Similarly, the work by Crane and D. Sornette also

[5]http://digg.com
[6]http://en.wikipedia.org/wiki/Part-of-speech_tagging
[7]http://svmlight.joachims.org/
[8]http://bitly.com
[9]http://www.liwc.net/

showed that quality of a Youtube video is strongly correlated with the way its rate of view counts varies overtime [7]. They further studied the probem of classifying Youtube videos based on their pattern of viewership. They fit the view counts time series after the point of the largest peak by the power law process. Then, videos are classified as *viral*, *quality*, or *junk* video based on the exponents of the power law.

In the social influence study, virality may shed more light in understanding user behaviors. Leskovec *et. al.* conducted an empirical study to understand effects of a viral marketing strategy in a large user-user recommendation network over time. They found that only a small fraction of customers choose to propagate product after purchasing and the deeper a user in the recomendation sequence, if they choose to forward the recommendation, the more people she tends to forward the recommendation to, subsequently, a large fraction of products (30%) have only one recommendation while top 10% products takes 84% of all recommendations. This means not all product become viral under the strategy. They also found that the number of purchases and the number of recommendations made by a single user and the size of recommendation cascade follow the power law distribution; the probability of purchasing a product increases with the number of recommendations received, but quickly saturates to a constant and relatively low probability; and probability that a user successfully recommend the product to a friend decreases with the increasing of the number of recommendations that the user sends to the friend. By identifying communities within user network using modularity based clustering [19], they observed that only very few products enjoy active recommendation within small communities.

VI. Conclusion

Virality modeling and analysis aim at quantitatively measuring the virality of items and individuals and how virality relates with other behaviors in large scale user-to-user networks. In this paper, we surveyed existing models, which mostly focus on one or more single facet of virality, measures each of facet based on some simple statistics. We presented the recent empirical studies showing the relationships between the virality, behaviors of user network and content. These correlationships have been applied to predicting popularity and decay time of items, based on other behaviors of early access to the items, and identifying astroturf content from virality.

In the future, we would like to model the virality of items and users in a common framework with regard to other behaviors. We want to further investigate effects of the underlying user network on the virality of different type of items, e.g. tweets, memes, production mentions, topics; and examine how virality correlates with other user behaviors, e.g. influence, and passivity.

Acknowledgement

This project was carried out at the Living Analytics Research Centre sponsored and supported by the Singapore National Research Foundation & Interactive & Digital Media Program Office, Media Development Authority.

References

[1] F. M. Bass. A new product growth for model consumer durables. *Manage. Sci.*, 50:1825–1832, December 2004.

[2] J. A. Berger and K. L. Milkman. Social Transmission, Emotion, and the Virality of Online Content. *Social Science Research Network Working Paper Series*, Dec. 2009.

[3] J. Bollen, H. Mao, , and A. Pepe. Determining the public mood state by analysis of microblogging posts. In *Proc. of the Alife XII Conf. MIT Press*, 2010.

[4] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. *Data Mining Workshops, International Conference on*, 0:296–304, 2010.

[5] R. Cowan. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, 28(8):1557–1575, June 2004.

[6] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[7] R. Crane and D. Sornette. Viral, Quality, and Junk Videos on YouTube: Separating Content From Noise in an Information-Rich Environment.

[8] R. Ferguson. Word of mouth and viral marketing: taking the temperature of the hottest trends in marketing. *Journal of Consumer Marketing*, 25(3):179.

[9] M. Guerini, C. Strapparava, and G. zbal. Exploring text virality in social networks. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[10] S. Helm. Viral marketing establishing customer relationships by "word-of- mouse". *Electronic Markets*, 10(3):158.

[11] T.-A. Hoang, E.-P. Lim, P. Achananuparp, J. Jiang, and F. Zhu. Virality of content in twitter: The experience in singapore's 2011 general election. In *Proceedings of the International Conference on Asia-Pacific Digital Libraries ICDL2011*, 2011.

[12] S. Jurvetson. From the ground floor: What exactly is viral marketing? *Red Herring Communications*, page 110, 2000.

[13] S. Jurvetson and T. Draper. Viral marketing. *Available from: http://www.dfj.com/news/article_26.shtml (accessed on Aug 9th 2011)*.

[14] J. Kirby and P. Marsden. *Connected Marketing: the Viral, Buzz and Word-of-mouth Revolution*. Oxford, UK: Butterworth-Heinemann, 2006.

[15] J. Lee, J.-H. Lee, and D. Lee. Impacts of Tie Characteristics on Online Viral Diffusion. *Communications of the Association for Information Systems*, 24(1), 2009.

[16] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1, May 2007.

[17] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[18] F. M. Modzelewski. Finding a cure for viral marketing. *Direct Marketing News*, 11, 2000.

[19] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, pages 8577–8582, 2006.

[20] M. Pastore. The value of word of mouth. *http://adres.internet.com/feature/article/ /0,1401,8961_395371,00.htm (accessed on Aug 9th 2011)*.

[21] S. Petrovi, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[22] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[23] E. M. Rogers. *Diffusion of Innovation*. The Free Press, New Yory, 1962.

[24] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[25] M. Rowe. Forecasting audience increase on youtube. In *Workshop on User Profile Data on the Social Semantic Web, 8th Extended Semantic Web Conference 2011 (ESWC 2011)*, 2011.

[26] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[27] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53:80–88, August 2010.

[28] S. Wu, C. Tan, J. Kleinberg, and M. Macy. Does bad news go away faster? In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.