International Journal of Advances in Software Engineering & Research Methodology– IJSERM
Volume 1 : Issue 2

Publication Date : 25 June 2014

Spread of Influence on Information Propagation: Target Viral Marketing through Influential Nodes in Social Network

Windelle John G. Vega
Luisa B. Aquino, MIT

Abstract*—Social Networking sites are already widely used nowadays in different aspects most especially on communication, information dissemination, marketing and others. In Social Network Analysis, one of the most studied problems is the information dissemination, and is often related to target viral marketing. The aim of this study is to provide a framework that consolidates the use of NodeXL as a tool for analyzing various data, determining the key influential nodes using the different metrics and to simulate the diffusion process using the two models (ICM and LTM). We have to first gather data from a Facebook network, do some analysis using NodeXL after which we will measure the metrics of each node to determine the key influential node over the network and lastly we will simulate the diffusion process using Linear Threshold Modeling and Independent Cascade Modeling. The results shall be the basis of the effectiveness of the proposed framework on Target Viral Marketing and it will show that the network structure has a great impact on the diffusion process. *(Abstract)

***Keywords—social media, influence propagation, information propagation, viral marketing, data mining* (key words).**

I. Introduction

The term Social Media, as defined by [8], is a graph of relationships and interactions within the groups of individuals. Social Networking Sites (SNS) are very popular because SNS are fast and free way of communication, it also has features that allow social interaction and interest sharing [7] and information can also be spread virally over social media through the process called Word of Mouth [5] [6].

According to [10] influential nodes on a Social Network (SN) can be used for the viral spread of information and this was supported in the study of [4]. [3] stated that identifying influential people is the first step in viral spread of information. These statements were interrogated by [1] upon saying that finding the influential people is challenging and was further argued upon by [8] in his question, how could we choose few key individuals to be seeds for the viral spread of information?

This paper aims to provide a study on how to maximize viral spread of information on SN through the influential nodes and apply this target viral marketing by introducing a framework that consolidates the use of social network graphing, clustering, influence mining and information dissemination modeling through Independent Cascade Model (ICM) and Linear Threshold Model (LTM) in analyzing information diffusion. This framework would incorporate the use of NodeXL as a tool. NodeXL is an MS Excel plug in that can be used to analyze different data over the Social Media. To the best of our knowledge, no studies have integrated the use of the said metrics yet. Previous studies were just focused on studying the influence of different nodes on SN while some focused on Data Mining Approach and some have just done Information Diffusion Modeling through ICM and LTM.

A. *Literature Review*

To clearly visualize different methods used by previous studies, the Table 1 below shows the different authors or studies with their corresponding methodologies and findings.

TABLE I. METHODS AND FINDINGS FROM PREVIOUS STUDIES

Author	Method	Method
Richardson and Domingos (2002)	Data mining for improved viral marketing.	Results show robustness and utility of the approach.
Kempe et. al (2003)	Influence Maximization using Independent Cascade Model and Linear Threshold Model.	The study show that a natural greedy strategy obtains solution that is probably within 63% of optimal for several cases of models.
Tang et. al (2009)	Topical Affinity Approach (TAA) to model the topic-level social influence on large networks.	The proposed approach can effectively discover the topical based social influences.

Windelle John G. Vega
University of Saint Louis
Philippines

Luisa B. Aquino, MIT
University of Saint Louis
Philippines

Author	Method	Method
Karmaker, et. al (2010)	Data Mining and Machine Learning Approach to generate efficient rule based decision for viral marketing	The classifier identifies the best ways to promote a particular game on Facebook depending on different attributes (country, game type, age, sex, credit card type). This achieved up to 78% accuracy.
Bonchi (2011)	Influence Maximization using Independent Cascade Model and Linear Threshold Model. Data Mining Perspective	Mainly focused on problems in influence maximization for viral marketing. Which model (ICM or LTM) does better describe the real world?
Gui-sheng et al (2011)	Attacks problem on selecting fixed number of initial users to maximize profits by implementing intelligent algorithms such as GA, DE, PSO	The model designed for solving viral marketing problem outperforms other current search methods.
Akrouf et. al (2013)	NodeXL to extract structure and type of relationship. Linear Threshold Model, Independent Cascade Model, Weighted Cascade Model (WCM).	Experiment show that the structure of the network affect the diffusion process directly.
Sharma and Shrivastava (2013)	Clustering application which is "Collaborative Data Mining" which mines different subgroups from some network in the form of graph such as SN.	Mining nodes with high connectedness, which in this study are considered influential, can be used for viral marketing or target advertising

As presented on the table above, previous studies regarding information diffusion over the social media proposed some sort of methods. Studies that have used Data Mining Approach can be seen from the works of [9] [7] [10]. In their study, data were gathered and analyzed and results showed how different information can spread over the network. Their approach considered different metrics which can be considered a very broad approach. Other studies, on the other hand, focused on Information Dissemination Modeling like ICM, LTM, WCM and TAA, [8] [11] [3] [2] and they are to select initial set of active nodes and use them as seeds on the different models and simulate how influence will spread over the network but on these studies they only assume which nodes to select as initial key nodes, no other metrics that were used. There were no explicit problems regarding the methods presented by these studies but if you come to think of it, it will be more efficient to combine these different process and form a much more efficient process or method on information dissemination over the social network. Thus, this study aims to combine some of the methods from previous studies to make the processes faster.

This study assumes that the use of influence mining and diffusion modeling will present a better picture of the path of diffusion. The influence mining utilizes the diffusion modeling to simulate the efficiency and effectiveness of a node if used as a seed of information diffusion and diffusion modeling as well utilizes influence mining to determine the initial seed for the diffusion as not to use random nodes as seeds. If we are to determine first, which of the nodes on the network are the most influential using different metrics such as Degreed Centrality, Betweenness Centrality, Closeness Centrality and Eigenvector Centrality, it is easier for us to model and simulate the spread of information and influence on the social media thus using the influential nodes as the seeds and making them as the initial set of active nodes on the dissemination modeling.

II. Methods

In this paper we have applied and tested the processes on two different networks to show that the process to be made won't just be applicable to a single kind of network. The first network shall be identified as Network A and the second network shall be identified as Network B.

Social Network Graphing. This will be the first method that will be used in order to achieve the goals of the study. In this method, the study will be able to form the graph of the social network, connect the different nodes and provide the weights to all the edges. This study shall make use of an existing tool called NodeXL to create the network graph and layout the graph using Fruchterman-Reingold Algorithm.

NodeXL. This study utilized NodeXL for the analysis of different data gathered. From a raw Facebook data, using the said tool, the study will be able to present these data to a clearer figures, both graphical and numerical for easier manipulation.

Clustering. As the next step of the experiment on influence propagation, the study will cluster each node using again NodeXL. In this way, the study could organize different nodes into clusters and can be able to determine the actors/nodes which have the common grounds and preferences known as strong ties in which they belong to strong cluster and those connected but belongs to different groups known as weak ties and they belong to weak cluster.

Influence Mining. In this method, the study shall find the different nodes which have very strong influence to other nodes by analyzing its links, social interactions on the social network and number of friends. These nodes will be the targets and they shall act as seeds for the viral spread of information. To perform this method these different metrics were calculated:

- *Degree Centrality* - is a simple count of the number of connections for each node.

- *Betweenness Centrality* - essentially reveals how important each node is in providing a "bridge" between different parts of the network.
- *Closeness Centrality* - is a measure of how close each node is, on average, to all of the other nodes in a network.
- *Eigenvector Centrality* - accounts not only for the node's own degree, the also the degrees of the nodes to which it connects.

Information Dissemination Modeling. To clearly visualize spread of information, this study would also show the influence diffusion or how the influential nodes or seeds could influence the other by using two widely and commonly used diffusion models which are the Linear Cascade Model and Independent Cascade Model.

- *Independent Cascade Model* - In the IC model, when a node v first becomes active, say at time t, it is considered contagious. It has one chance of influencing each inactive neighbor u with probability pv,u, independently of the history thus far. If the tentative succeeds, u becomes active at time $t+1$. The probability pv,u, that can be considered as the strength of the influence of v over u.
- *Linear Threshold Model* - In the LT model, each node u is influenced by each neighbor v according to a weight pv,u, such that the sum of incoming weights to u is no more than 1. Each node u chooses a threshold θu uniformly at random from $[0, 1]$. At any timestamp t, if the total weight from the active neighbors of u is at least θu, then u becomes active at timestamp $t + 1$.

The bottom line is that, if we find first the key influential nodes, it would be easier to model how information shall propagate such that the key influential nodes will become the initial set of active nodes on the dissemination modeling.

III. Results and Discussions

Graph representation of a social network is an apt abstraction of their connected nature and allows us to tap into the rich repository of graph and complex network theories (Chun, Kwak, Eom, 2008). This section shows the network graph using the metrics of Social Network Graphing, Clustering and Information Mining based on Degree and Betweenness Centrality. The network graph shall be used to simulate the spread of information and influence over social media. Consequently, influential nodes will be best determined and making them as the initial set of active nodes on the dissemination modeling.

A. *Social Network Graphing*

The network graphs as shown in Figure 1, through the use of the Fruchterman-Reingold algorithm created a visual illustration for easier detection of various parts.

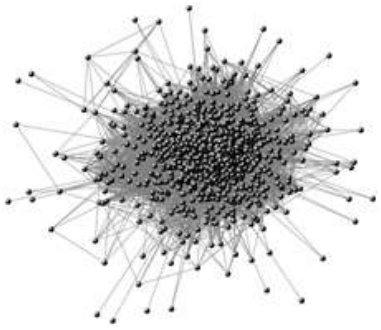

Figure 1. Graph for the Social Network Friends

B. *Clustering*

Figure 2 show the graphical illustration of how the nodes in the network are logically grouped using Clustering. The clusters indicate which nodes are the closest and which are the farthest on a certain organization.

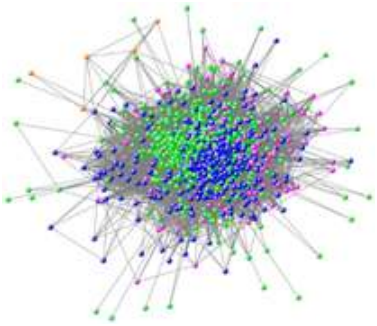

Figure 2. Clustered Network using Clauset-Newman-Moore Algorithm

C. *Influence Mining*

Figure 3 show the results when the different metrics used were computed. The figure only show the top 10 nodes sorted by Degree, Betweenness Centrality, Closeness Centrality and Eigenvector Centrality.

Graph Metrics

Degree	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality
414	8511.937	0.001	0.004
393	10246.908	0.001	0.004
389	15961.487	0.001	0.003
372	4949.050	0.001	0.004
360	7882.787	0.001	0.002
345	6397.872	0.001	0.003
343	11777.957	0.001	0.003
333	7805.841	0.001	0.002
324	3477.764	0.001	0.004
321	4097.795	0.001	0.004
320	4571.467	0.001	0.004

Figure 3. Calculated Metrics of the Nodes

Figures 4-7 show the influential nodes in each cluster based on the different metrics presented. The level of influence of each node are identified through their sizes. The bigger the node is, the more influential it is.

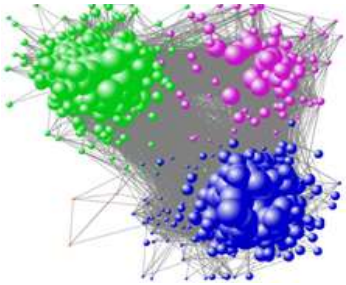

Figure 4. Influence Based on Degree

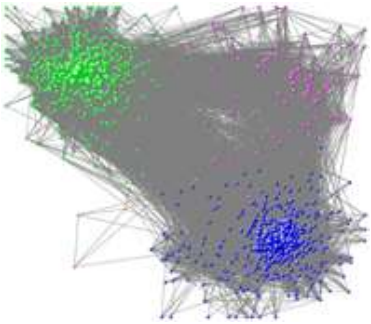

Figure 5. Influence Based on Betweeness Centrality

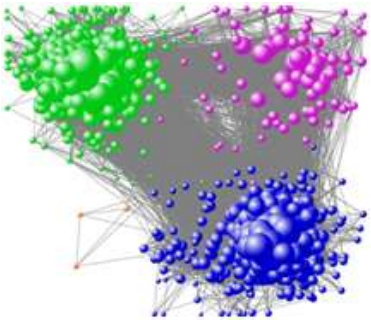

Figure 6. Influence Based on Closeness Centrality

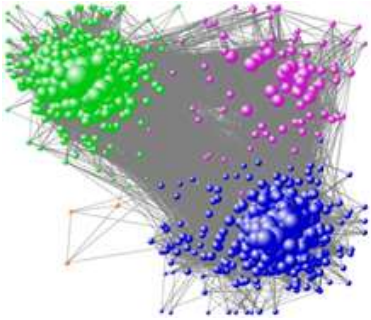

Figure 7. Influence Based on Eigenvector Centrality

On the question thrown by [8] about how could we choose few key individuals to be the seeds on information diffusion? This paper argues that the most influential on the network can be identified using the Social Network Graphing, Clustering and Influence Mining. These metrics were able to show clearly the influential nodes, thus these nodes will be used as the seeds of the viral spread of information, as supported by the study of [10].

D. *Diffusion Modeling*

The diffusion modeling created a simulation on how the information will be spread all throughout the network. This study simulated the information diffusion using random nodes as seeds and using the predetermined influential nodes as seeds then compared it to see if there's a gap if different nodes are used as initial paths of diffusion.

E. *The Framework*

Figure 8 illustrates the structure of the proposed framework in this study. This shows how the raw social network data turns into useful information regarding the target viral marketing strategy. The data will undergo several layers and as it go beyond; some sort of useful information will be extracted as presented under the various methods that were done in this study.

The first layer illustrates the Facebook raw data gathered from the selected social network. These data are those which will be used on the process for further analysis. From the Facebook raw data, as it will go to the next layer, the graphical illustration of its graph will be presented. Then the next layer will divide the network graph into various groups or clusters. This will now define which cluster are the strong clusters and which cluster are the weak clusters. On the next layer, the influential nodes in each cluster will be identified to break it down into smaller pieces. The next layer will now simulate the dissemination process or how will the marketing information will be spread all throughout the network using the dissemination modeling. Upon reaching the last layer, the useful information will now be produced and these will be used for a more efficient dissemination of information and marketing strategy.

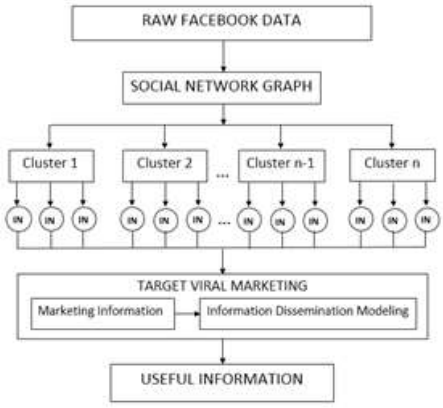

Figure 8. The Proposed Framework

IV. Conclusion

The findings of this study showed promising results. On the influence mining, after calculating the difference metrics of each node, it showed that almost the same nodes have the strong influence in each metric. In other words, although there

were different metrics that had been utilized on influence mining, the same nodes will still be identified as influential but the probability that each node can influence others will still differ from one metric to another considering other factors. After the simulation, it was observed that the diffusion process was greatly affected by the network structure, the same finding was observed on the study of [2]. Compared to the process wherein the seeds are chosen randomly, there is just a small difference in the propagation if the key influential nodes will be considered as seeds. But as the network becomes larger, the difference will also increase. With that premise, it can be concluded that the proposed framework can be applied on large network.

v. Future Works

This study suggests that instead of using different metrics for the influence mining, since it was found out that almost the same nodes will be identified as influential if different metrics are used, they may just use one metric or if it is possible they may still compute the different metrics of each node and combine the different results to come up with a summarized computation. Future studies may also use just the Independent Cascade Model for the simulation of the diffusion process since each node can give different probabilities of activation unlike the Linear Threshold Model. As a general suggestion, future studies may introduce a framework that works on both small scale and large scale networks and that it can identify influential nodes as seeds on information diffusion.

References

[1] Abedniya, A., Mahmouei, S.S. The Impact of Social Networking Websites to Facilitate the Effectiveness of Viral Marketing. International Journal of Advanced Computer Science and Applications, Vol. 1, No. 6, 2010.

[2] Akrouf, S., Meriem, L., Yahia, B., Eddine, M.N. Social Network Analysis and Information Propagation: A Case Study Using Flickr and YouTube Networks. International Journal of Future Computer and Communication, Vol. 2, No. 3, 2013.

[3] 5Bonchi, F. Influence Propagation in Social Networks: A Data Mining Perspective. IEEE Intelligent Informatics Bulletin, Vol. 12, No. 1, 2011.

[4] 8Hsu, W., Lancaster, J., Paradesi, M., Weninger, T. Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach. ICWSM'2007 Boulder, Colorado, USA, 2007.

[5] 9Hu, H.-W., Lee, S.-Y. Study on Influence Diffusion in Social Network. International Journal of Computer Science and Electronics Engineering, Vol. 1, Issue 2. ISSN 2320- 4028, 2013.

[6] 10Jankowski, J., Ciuberek, S., Zbieg, A. Studying Paths of Participation in Viral Diffusion Process. SocInfo 2012, LNCS 7710, Springer-Verlag Berlin Heidelberg, 2012.

[7] 12Karmaker, D., Rahman, H., Rahaman, M.S., Bari, K. A Fine Grained Technique for Viral Marketing based on Social Network: A Machine Learning Approach. International Journal of Science and Technology, Vol. 1, No. 2, 2011.

[8] 13Kempe, D. et al. Maximizing the Spread of Influence through a Social Network. Association of Computing Machineries, 2003.

[9] 15Richardson, M., Domingos, P. Mining Knowledge-Sharing Sites for Viral Marketing. Association of Computing Machineries, 2002.

[10] 17Sharma, S., Shrivastava, V. Viral Marketing in Social Network Using Data Mining. International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 1, Issue 4, 2013.

[11] 19Tang, J., Sun, J., Wang, C., Yang, Z. Social Influence Analysis in Large-scale Networks. Association of Computing Machineries, 2009.

About Author (s):

Windelle John G. Vega
Student of Bachelor of Science in Computer Science at University of Saint Louis, Tuguegarao City
Currently living at Blk. F, Lot 17, Woodcrest Subdivision, Alimannao, Peñablanca, Cagayan, Philippines.

Luisa Baquiran Aquino is the Academic Dean of the ITE programs of the University of Saint Louis (USL), Tuguegarao City, Cagayan, Philippines. Her research interests cover software engineering, project management, e-learning practice and studies related to educational and curriculum development.