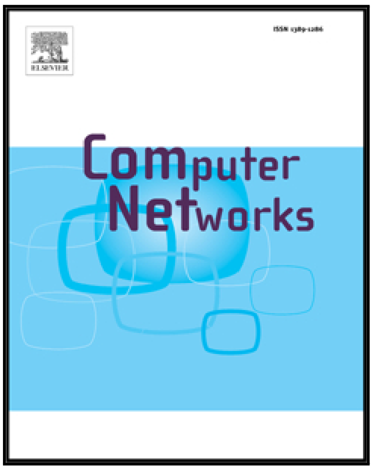
Journal Pre-proof



Analysis and Evaluation of Random-Based Message Propagation Models on the Social Networks

Wei Kuang Lai , Yi Uan Chen , Tin-Yu Wu

PII: S1389-1286(19)30560-2
DOI: https://doi.org/10.1016/j.comnet.2019.107047
Reference: COMPNW 107047

To appear in: *Computer Networks*

Received date: 5 May 2019
Revised date: 16 November 2019
Accepted date: 7 December 2019

Please cite this article as: Wei Kuang Lai , Yi Uan Chen , Tin-Yu Wu , Analysis and Evaluation of Random-Based Message Propagation Models on the Social Networks, *Computer Networks* (2019), doi: https://doi.org/10.1016/j.comnet.2019.107047

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier B.V.

Journal Pre-proof

Analysis and Evaluation of Random-Based Message Propagation Models on the Social Networks

Wei Kuang Lai, *Senior Member, IEEE*, Yi Uan Chen, Tin-Yu Wu*

***Abstract*—Social network services (SNS), based on the complex relationships among people in real-life and virtual world, have become a major internet service for people to communicate with each other. Different social networks have different characteristics and varying levels of influence. To understand the message propagation process, the driving power behind it and its social influence, this paper presents a detailed analysis of message propagation models over the social networks by analyzing the relationships among nodes. This paper presents five proposed models which aim to analyze message propagations on social networks. We analyze the message propagation models and show how messages spread through the social networks. Furthermore, we propose a social network analysis on Hadoop platform to verify the social network characteristics. We also present a measurement study of messages collected from 900K users on Facebook, to verify our proposed models by means of big-data Hadoop platform. We believe that our research provides valuable insights for future social network service research.**

***Index Terms*—SNS (Social Network Service), MPM (Message Propagation Model), Social Cluster**

Journal Pre-proof

I. Introduction

Recently, the rapid development and advancement of social network service (SNS) has made it possible to connect people who share the same interests and activities across political, economic, and geographic borders. Social network services have become a major internet service for people to communicate with each other. The current social network services generally operate based on the Six Degrees of Separation [1], which suggests that two random people are able to connect by a chain of six acquaintances on average, and aims to help users expand their personal networks through friends and connections. The mathematical analyses of the social network services revealed that most of the social network data and structures are too large and complex to be transformed into strict mathematical description. Therefore, computer simulations or big data analyses has become an accredited scientific verification, but how to collect useful information is a topic worth discussing. Nowadays, many researchers are interested in how to discover the rule or structure behind the complex social networks.

In this paper, we present five proposed models to analyze message propagation on social networks. We analyze the message propagation models and show how messages spread through the social network. The mathematical models for the message propagation on social networks are proposed. The MPM message propagation model is constructed based on the social network message propagation process, and the features of social communities. We present a measurement study of messages collected from 900K users on Facebook, to verify and modify our proposed model by means of big-data Hadoop platform. Different social influences are evaluated by examining their stabilities, evaluations, correlations and clusters. [24] [25]

The rest of this paper is organized as follows: Section II reviews researches on social network background, structures of social networks and influence factors of social relations. Section III details how we adopt message propagation analyses to build our models. We first define the simple RM model, two kinds of roles, and methods of how to find their relations, and then we propose RMR model, RMO model, RMI model and revised model RLHMl. Step by step, we take more details into consideration and adjust our models to match the real-life conditions on social networks. In Section IV, we perform the experiment and simulate the parameters with our proposed models. Finally, we summarize our findings and results and suggest some other problems and applications to be addressed in the future.

II. Socially Influential Propagation Models

Social influence is essentially dynamic. In social activities, everyone's influence in social groups not only varies with his words, deeds and social characteristics, but also propagates with diverse social activities on social networks. Therefore, analyses and research on the dynamic propagation process of social influence is of great significance in understanding the natures and characteristics of social influence, in realizing the formation and evolution of social networks, and in discovering the rules of information propagation and people's behavior patterns and many other issues on social networks. In the classic propagation models [2] proposed by Katz and Lazarsfelds, the propagation of information or innovations started from groups with strong social influence, and then via them further spread to a wider range of crowds. The most socially influential propagation problems were first proposed by Richardson and Domingos [3, 4]. In simple terms, the most socially influential propagation problems mean the search for the most influential members on social networks. Kempe, Kleinberg and Tardos [5, 6] formalized the issue, and

Journal Pre-proof

concluded an Independent Cascade Model, a Linear Threshold Model and the generalized model of the two models. Then they utilized the discrete optimization methods to solve the problem. In comparison, Independent Cascade Model was more widely applied in many missions while Linear Threshold Model was preferred when only the cumulative effects which the neighbors' influence would have on users were taken into consideration. In the above models, the default value of the influence weights between individuals is a known constant or a numeric value given in advance, but this assumption is not true in the real environment. Thus, a lot of research has analyzed and quantified the factors of individual influence and their interrelationships. Goyal, Bonchi and Lakshmanan [7] later utilized the user action log to measure the influence coefficient and achieved better results. There was also some research attempting to use the network topology and the historical interactive information, in the case of missing log data, to quantify the influence among users [8]. Because the most influential propagation problem is NP-hard [9], the heuristic algorithms for the problem have been profoundly studied.

(a) The influence propagation model.

Social influence takes effect through people's daily interactions. On social networks, the main interaction is through publishing, sharing and propagating information; therefore, the working process of influence on social networks and the information propagation process have an inherently close connection and a very similar mechanism. The information propagation model thus plays a very important role in the research of influence propagation. Guille, Hacid, Favre and Zighed [10] present a more detailed introduction about message propagation.

(b) Independent Cascade Model (ICM) [11-14]

The fundamental principles of Independent Cascade model are similar to those of SIR (Susceptible Infective Recovered) model, which is utilized to describe the spreading of infectious diseases [15]. ICM can be described as: on social networks $G = (V, E)$, part of users $V_0 \in V$ are in the start state at initial moments, and the influence between user v_i and its neighbor v_j can be represented by $p_{i,j}$. The take value of $p_{i,j}$ is independent; it would not be influenced by the relationship between v_i and its neighboring nodes in the propagation process. If user v_i initiates neighboring nodes from a non-start state into the start state at a certain moment *t*, user v_i has only one chance to try to activate each neighboring node at the *t* moment. For example, when the user v_i's neighbor, v_j, is in non-start state at the *t* moment; v_i would activate v_j with probability $p_{i,j}$. If the activation is successful, v_j would be in the start state from the *t+1* moment on. But whether it is successful or not, v_i can no longer try to activate v_j. If multiple v_j's neighbors are in active state at the t moment, the order in which they try to activate v_j is arbitrary. The system would start the propagation process from the initial state until no more new users are activated.

(c) Linear Threshold Model (LTM) [15-17]

Linear Threshold Model is widely used in the adoption of new products and it is the core of many threshold models [14]. This model can be described as: on social networks $G = (V, E)$, the influence weight between user v_i and its neighboring node, v_j, is $W_{ij,}$ and the maximum combined influence weight of all the v_i's neighboring nodes is 1; that is,

$$\sum_{v_j \in Ng_i} w_{i,j} \leq 1 \qquad (1)$$

any user, v_i, would randomly select their own threshold values $\theta_i \in [0,1]$, which means that only when the influence which v_i's neighbor has on v_i exceeds a threshold will v_i be activated. Similar to Independent Cascade Model, the collection of users that stay in the start state at the initial moment is defined as V_0. If the collection of v_i's neighbors who are activated at the *t* moment is Ng_i^{act}, and v_i has yet to be activated, v_i can be activated as follows:

$$\sum_{v_j \in Ng_i^{act}} w_{i,j} \geq \theta_i \qquad (2)$$

v_i turns into the start state from the *t+1* moment on, and keeps remaining in that status. Each node's tendency to become active would increase monotonically as more of its neighbors become active. The system would keep activating nodes from the initial state until no more new users are activated.

(d) Expansion and other models

Kempe and Kleinberg and Tardos [18] conducted expansion on ICM and LTM, and modified the independence conditions [14] in the models. In order to modify an Independent Cascade Model, they defined incremental function $p_{vj}\ (v_i, S_j)$, which represents the probability of user v_i activating v_j, and S_j was the collection of neighboring nodes which attempted to activate v_j but eventually failed. To modify the Linear Threshold Model, they defined a value function $f_{vi}\ (Y_i)$, in which Y_i was the collection of user v_i's neighbors which had been activated at a previous moment, and then the conditions of node v_i getting activated became:

$$f_{v_i}(Y_i) \geq \theta_i \qquad (3)$$

The above two expansion models are equivalent to each other in framework, so that they can be easily converted. They found that the probability of user v_i being activated by neighbors decreased over time; that is, the more neighbors of user v_i tried to boot v_i without success, the less impact the newly activated neighbors would have on user v_i, namely:

$$p_{v_i}(v_j, S_i) \geq p_{v_i}(v_j, S_i'), S_i \subseteq S_i' \qquad (4)$$

Based on the above findings, Decreasing Cascade Model was devised for the modeling of target selection. Bass [19] proposed a product diffusion model, which can also be used to describe the propagation process of social influence. The propagation of influence in the model was in s-shaped distribution: the propagation among users was initially slow, then suddenly increased exponentially and finally became slow again. This model was used more often on qualitative analyses rather than on quantitative calculation.

Discussion about the socially influential propagation problems: The widely-applied ICM can properly describe the spread of an epidemic, as well as the information propagation on networks. But from the angle of social influence, many ICM theories are contradictory to real facts: users in ICM will only be activated by a certain neighbor; namely, users on social networks are most likely to have resonance with a certain socially influential neighbor, and the influence of other neighbors will be eventually ignored; user influence will only take effect one time on neighbors in the propagation process, and the effect will occur simultaneously. In both ICM and LTM,

Journal Pre-proof

users would generate some changes in status after being affected by some influence; the changes are irreversible, and it is clearly inconsistent with the reality. Therefore, it would be challenging to devise message propagation models that are more in line with the characteristics of influence on social networks.

III. Message Propagation Model (MPM) for Social Networks

The influence also spreads with social activities on SNS. Therefore, analyses and research on the dynamic propagation process of social messages are greatly significant in understanding the nature and the characteristics of the impact of the social network, in realizing the formation and evolution of the social network, and in discovering the message propagation and people's behavior patterns on the social networks.

As mentioned in Section II, ICM model can effectively describe the problems, such as the social message propagation and the information diffusion on the social networks. But if we examine from the viewpoint of social networks influence, many viewpoints about ICM model are found to go against facts. For example, in ICM model, users will only be activated by certain neighbors; that is, the social networks users are most likely to resonate with certain neighbors with strong influence and other neighbors' influence will eventually be ignored. Besides the limitation, during the influence propagation process, the user influence will only take effect once on its neighbors simultaneously. In both ICM and LTM models, users would experience certain changes in status after being affected by certain influence and the status change is not reversible, which is clearly inconsistent with the real-life situation. Therefore, we hope to devise a MPM message propagation model, which is more in line with the characteristics of the social network influence.

Since we take too many factors into account, the computing cost would be high because it is not easy to solve an NP-Hard problem. Due to the increasingly expanding social network, designing an uncomplicated message propagation model, which is capable of predicting the social networks message propagation trend and enhancing the prediction efficiency, is the main research objective of this paper. The message propagation model can be applied to describe the process of message propagation on the social networks and further to analyze the regularity of the transformation of message users from receivers to messengers. The characteristic of the social network service, propagating messages right after receiving them, enables us to predict the peak time of message propagation and in turn to achieve the efficient message propagation. The construction approach of the message propagation models is mainly based on the analysis of the message propagation evolution on the social network service. Five message propagation models are proposed in this study, including RM (receiver-messenger) Model, RMR (receiver-messenger-receiver) Model, RMO (receiver-messenger-outer) Model, RMI (receiver-messenger-idled) Model and RLHmI (Receiver-Light-Heavy-Messenger-idled).

A. *RM Model*

On account of the complexity of the social network service, it is difficult to mathematically analyze the message propagation. Thus in this study, we will rely primarily on simulations to verify mathematical analyses. In this way, we can focus on realistic scenarios of message propagation on the social network service.

Assumption 1-Small world effect: SNS, like Facebook, has the small-world phenomenon.

Assumption 2- Cluster coefficient: In SNS, users with a larger number of friends, which represents a higher cluster coefficient, can make new friends much easily through their original friends and their messages will be propagated much more rapidly.

Assumption 3- The time span online: In SNS, the longer timespan the users spend online, the more easily they will receive friends' messages and the more likely the messages will be propagated to other users.

Assumption 4- The individual and environmental separation degree: In SNS, due to the influence of environments or privacy and even because of some companies' strict management, users may reduce the timespan spent online and eventually become outers of SNS, which would definitely decrease the message propagation efficiency.

Let us first describe the hypothesis of message propagation scenario in the first RM model: We assume that every node has an equal chance, per unit time, of coming into contact with every other node at random in this model. Our RM Model is established on the condition that messages are propagated in the closed social networks and that the total Facebook population remains the same. The immigration rate (new users) and the emigration rate (dropouts) are not considered. Once users have contacts with messages, they would become messengers and cannot turn back into receivers in Figure 1.

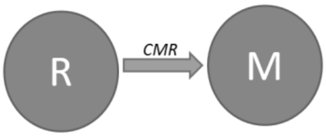


Figure 1. RM model diagram

First, we will examine the message propagation through the nodes on Facebook. Assume that *R(t)* is the number of nodes who are receivers at time *t*, and *M(t)* is the number of nodes who are messengers. Technically, due to the random process of message propagation, these numbers are not absolutely fixed— if the same messages spread on Facebook more than once, even under identical conditions, the numbers would probably vary each time. To tackle this problem, we define *R(t)* and *M(t)* more specifically to be the average of the expected number of receivers and messengers nodes, i.e. the average numbers we would get after we conduct the process many times under the same conditions. When receivers get the message from messengers, the number of messengers would thus increase. Providing that people on social networks make contacts to contribute to the propagation of the message randomly with a contact rate of *C* (cluster coefficient) in a time unit, that means each node has *C* contacts on average with randomly chosen others per time unit.

Only when a receiver node has a contact with a messenger node would the message be propagated. If the Facebook is composed of *N* nodes, the average probability that

Journal Pre-proof

the node you meet at random is a receiver would be *R(t)/N*. Thus, a messenger node has a contact with an average of *CR(t)/N* receiver node per time unit. Since there are a total of *M(t)* messenger nodes on average, it means that the average rate of new messengers will be *CR(t)M(t)/N* and we can propose a differential equation for the rate of change of *M(t)* as follows:

$$\frac{dM(t)}{dt} = C\frac{R(t)M(t)}{N}$$

It is often convenient to define variables representing the fractions of receiver and messenger nodes, thus:

$$\frac{R(t)}{N} = r(t), \qquad \frac{M(t)}{N} = m(t)$$

In terms of the above equation, the equation can be written as

$$\Rightarrow \frac{dm(t)}{dt} = Cr(t)m(t) \qquad (5)$$

Because $m(t) + r(t) = 1$, then

$$\frac{dm(t)}{dt} = Cm(t)\big(1 - m(t)\big) \qquad (6)$$

If the initial value of t is set as $t = 0$ and the initial value of messengers as m_0, we put them in Equation (6) and get $m(t)$.

Then, the solution $\Rightarrow\ m(t) = \dfrac{1}{1 + (\frac{1}{m_0} - 1)e^{-Ct}}$ (7)

Because $Cm(t)r(t)$ represents the increasing rate of messengers $m(t)$ and meanwhile the decreasing rate of message receivers $r(t)$, it can be proposed that $\frac{dr(t)}{dt} = -Cm(t)r(t)$.

Based on the above hypothesis, RM model can be represented by the following equations:

$$\begin{cases} \frac{dm(t)}{dt} = Cm(t)r(t) \\ \frac{dr(t)}{dt} = -Cm(t)r(t) \\ m(t) + r(t) = 1 \end{cases} \qquad (8)$$

In this RM model, the ratio of the numbers of messengers transforming into receivers can be illustrated in the following curve chart.

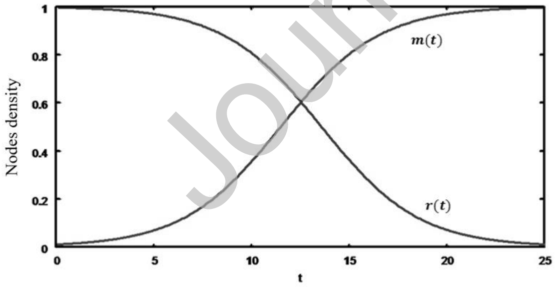


Figure 2. RM model curve chart

Based on Equation (7), when $t \to \infty$ and $m \to 1$, all the message users on the social networks will ultimately be changed into messengers, as illustrated in Figure 2. However, it definitely does not conform to real-life situations, in which people have contact C with only a small fraction of users on Facebook in the social network services, and that fraction is not chosen randomly. Of course, this cannot represent the real world because in this model, we ignore the possibility that messengers are likely to function as receivers as well or that message users may withdraw from social networks. To take the two above-mentioned variable factors into consideration, the RM model can be further modified into the RMR model.

B. RMR Model

The hypothesized scenario of RMR model is basically the same as that of the RM model shown in Figure 3, but we take one more factor into consideration: messengers are likely to turn back into message receivers for some reason in certain time unit. If there will be U percent of messengers changing into receivers in a time unit, U can be used to represent the off-line rate (the rate of M → R). Moreover, the receivers transforming from messengers may be the potential messengers sometime in the future. Therefore, $\frac{1}{U}$ can be interpreted as the average cycle of the message propagation. The changing rate of message users turning from messengers to receivers in a time unit would be denoted by $UM(t) = UNm(t)$.

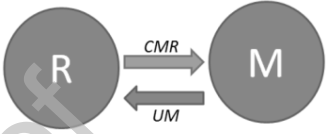


Figure 3. RMR model diagram

Based on the factor mentioned above, we can modify the RM model as follows.

Equation (5) will be modified as:

$$N\frac{dm(t)}{dt} = CNm(t)r(t) - UNm(t) \qquad (9)$$

If we subtract the decrease of the number in messengers from the increase of the number in messengers in the time unit, we can get the change of the number in messengers on the social networks. Therefore, Equation (8) can be modified as:

$$\begin{cases} \frac{dm(t)}{dt} = Cm(t)\big(1 - m(t)\big) - Um(t) = -Cm^2(t) + (C - U)m(t) \\ \qquad\qquad = \big(-Cm(t) + (C - U)\big)m(t) \\ m(0) = m_0 \end{cases}$$

In this way, we can get the solution in the stable condition as follows:

$$m(t) = \begin{cases} \dfrac{C - U}{\frac{C - Cm_0 - U}{m_0 e^{(C-U)t}} + C}, C \neq U \\ \left(Ct + \frac{1}{m_0}\right)^{-1}, C = U \end{cases} \qquad (10)$$

Definition 1: $\sigma = \frac{C}{U}$ (11)

C represents the contact rate of the cluster of messengers on social networks in a time unit and $\frac{1}{U}$ can be described as the average cycle of the message propagation. So, $\frac{C}{U}$ represents the relative rate of receivers to messengers and messengers to receivers.

For $m(0) \in (0,1)$, if we set $t \to \infty$ in Equation (10).

$$\lim_{t\to\infty} m(t) = \begin{cases} 0 & ,C < U \\ \left[\frac{C}{C-U} + 0\right]^{-1} & ,C > U \\ \left[\infty + \frac{1}{m_0}\right]^{-1} & ,C = U \end{cases} \Rightarrow \begin{cases} 0 & ,C < U \\ 1 - \frac{U}{C} & ,C \neq U \\ 0 & ,C = U \end{cases}$$

If we put formula (11) in the equation, then

$$m(t \to \infty) = \begin{cases} 1 - \frac{1}{\sigma}, \sigma > 1, and\ \frac{C}{U} > 1 \Rightarrow C > U \\ 0,\ \sigma \leq 1 \end{cases} \qquad (12)$$

The σ value is a key value on social networks because we can utilize this value to determine whether there will be efficient

Journal Pre-proof

Journal Pre-proof

message propagation.

Theorem 1:

When the parameter $\sigma \leq 1$, the range of the message propagation in this model will gradually shrink and the message propagation cycle will end soon.

Proof: from Equation (12), if $\sigma = \frac{C}{U} \leq 1$, in Figure 5, $m(t)$ will become smaller over time and gradually approach 0. In other words, the message propagation will eventually be in control. But when $\sigma > 1$, the model will be in a stable condition and$m^*(t) = \frac{C-U}{C} > 0$. Therefore, if $0 < m_0 < m^*(t)$, the messengers will gradually increase to $m^*(t)$, which means the expansion of the message propagation. But, if $m_0 > m^*(t)$, the messengers will gradually decrease to $m^*(t)$ and the message propagation will definitely shrink.

Theorem 1 reveals that if we want to expand the range of message propagation, we can set $\sigma = \frac{C}{U} \geq 1$. In Figure 4, as C is larger, U would be smaller. The larger C (clustering coeffient) represents those users who have a larger number of friends. A system cluster with a higher cluster coeffient would be able to make new friends much more easily through their original friends and the messages therefore will be propagated much more rapidly. $1/U$ denotes the average cycle of the message propagation, and a larger value signifies a longer duration of users' involvement in social activities.

To sum up, we can conclude this model equation as

$$\begin{cases} \frac{dm(t)}{dt} = Cm(t)r(t) - Um(t) \\ \frac{dr(t)}{dt} = -Cm(t)r(t) + Um(t) \\ r(t) + m(t) = 1 \end{cases} \qquad (13)$$

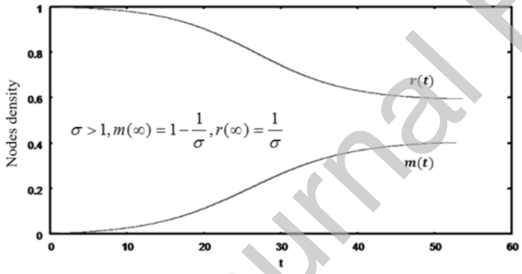


Figure 4. RMR model curve chart when $\sigma > 1$

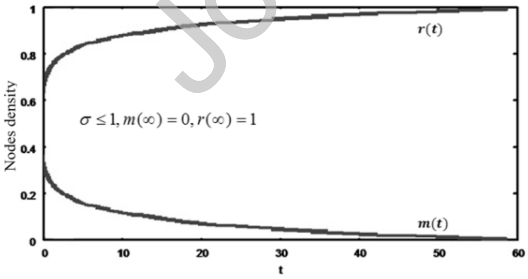


Figure 5. RMR model curve chart when $\sigma \leq 1$

The RMR model is more applicable to simple social network services, like blogs and social forums, which are lacking in user interactions.

C. RMO Model

In this model, message users on the social networks are also divided into messengers and receivers. But, we take one more factor into account: some message users may withdraw from the social network services and become an outer of the SNS, as shown in Figure 6.

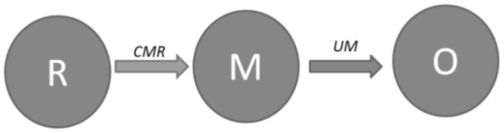


Figure 6. RMO model diagram

Therefore, in this model we add one more variable, *O(t)*, which represents the rate of the number of the outers in the SNS in the time unit *(t)*. Considering the new variable, equation $m(t) + r(t) = 1$ should be modified as $m(t) + r(t) + o(t) = 1$. $UNm(t)$ indicates the total number of the outers in the SNS in the time unit.

So, $N\frac{do(t)}{dt} = UNm(t)$ (14)

We set the initial value $r_0 > 0, m_0 > 0$ and consider that o_0 may be greater than 0 or equal to 0. $o_0 = 0$ is generally preferred. With Equations (9), (10) and $m(t) + r(t) + o(t) = 1$, we can get another equation set.

$$\begin{cases} \frac{dm(t)}{dt} = Cm(t)r(t) - Um(t) \\ \frac{dr(t)}{dt} = -Cm(t)r(t) \\ \frac{do(t)}{dt} = Um(t) \\ m(t) + r(t) + o(t) = 1 \\ m(0) = m_0, r(0) = r_0 \end{cases} \qquad (15)$$

Theorem 2: In RMO model, $r(t)$ will monotonically decrease to a non-zero number.

Proof: Because $m(t) \geq 0,\ r(t) \geq 0$ from Equation (15), we know $\frac{dr(t)}{dt} \leq 0$ and $r(t)$ will monotonically decrease. As $r(t)$ means the rate, we have $0 \leq r(t) \leq 1$. Therefore $r(t)$ is converged. From theorem 2, we know that $\lim_{t \to \infty} r(t) = r_\infty$ exists. In Equation (15), we divide $\frac{dm(t)}{dt}$ by $\frac{dr(t)}{dt}$, and get:

$$\frac{\frac{dm(t)}{dt}}{\frac{dr(t)}{dt}} = \frac{Cm(t)r(t) - Um(t)}{-Cm(t)r(t)} = -1 + \frac{U}{Cr(t)} = \frac{dm(t)}{dr(t)}$$

If we apply Equation (11) and get $\frac{dm(t)}{dr(t)} = -1 + \frac{1}{\sigma r(t)}$

The

$$\begin{cases} \frac{dm(t)}{dr(t)} = \frac{1}{\sigma r(t)} - 1, \Rightarrow \int_0^t \frac{dm(t)}{dr(t)} dr(t) = \int_0^t \left(\frac{1}{\sigma r(t)} - 1\right) dr(t) \\ m(t)|_{r=r_0} = m_0 \end{cases} \qquad (16)$$

The solution is:

$$m(t) = (r_0 + m_0) - r(t) + \frac{1}{\sigma}\ln(\frac{r(t)}{r_0}) \qquad (17)$$

The following is the analysis of the variation of *m(t)*, *r(t)*, and *o(t)*, based on Equations (15), (17). When (*t*→∞), their limit values are respectively m_∞, r_∞ and o_∞.

Theorem 3:

If $r_0 > \frac{1}{\sigma}$, $r(t)$ will increase and then decrease to $r = \frac{1}{\sigma}$. When $= \frac{1}{\sigma}$, m(t) will have the maximum value. As $m(t)$ tends to decrease, $r(t)$ will also monotonically decrease to r_∞.

If $r_0 < \frac{1}{\sigma}, m(t)$ tends to decrease, $r(t)$ will monotonically decrease to r_∞.

Proof: In Equation (17) $(m = (r_0 + m_0) - r + \frac{1}{\sigma}\ln(\frac{r}{r_0}))$, we

Journal Pre-proof

Journal Pre-proof

conduct a partial differential of $r(t)$, and get $\frac{dm(t)}{dr(t)} = -1 + \frac{1}{\sigma r(t)}$. When $\frac{dm(t)}{dr(t)} = 0$, m has an extreme value. So, when $r = \frac{1}{\sigma}$, the extreme value will exist. Then we can conduct a second partial differential of $r(t)$ to get $\frac{d^2m(t)}{dr^2(t)} = -\frac{1}{\sigma r^2(t)}$. Because $\sigma = \frac{C}{U} > 0$ and $r^2(t) > 0$, $\frac{d^2m(t)}{dr^2(t)} < 0$. When $= \frac{1}{\sigma}$, m will have the maximum value.

Theorem 2 shows that $r(t)$ will monotonically decrease to r_∞ and have low bound 0.

We know when $r = \frac{1}{\sigma}$, m will have the maximum value and $r(t)$ will monotonically decrease to r_∞. Thus, $m(t)$ decreases with the decreasing $r(t)$.

Equation (16) shows that when $t = 0$ and $r(0) = r_0 > \frac{1}{\sigma}$, $m(t)$ will increase over time to the maximum value of the development and later gradually decrease until it disappears eventually. This means that if $r_0 > \frac{1}{\sigma}$ and $r_0 \frac{C}{U} > 1$, the message propagation on the social networks will be very rapid and efficient in Figure 7. If $r_0 < \frac{1}{\sigma}$, *r(t)* and *m(t)* will decrease as time *t* increases in Figure 8.

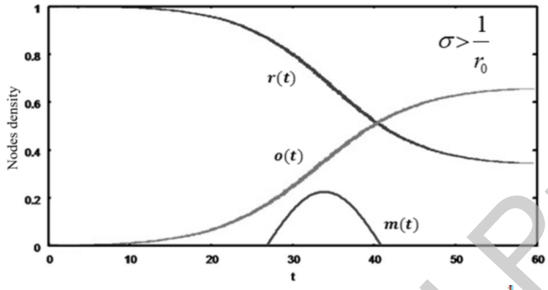


Figure 7. RMO model curve chart when $r_0 > \frac{1}{\sigma}$

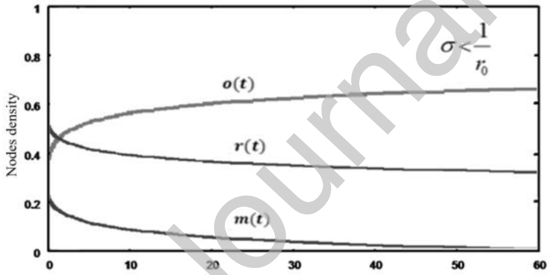


Figure 8. RMO model curve chart when $r_0 < \frac{1}{\sigma}$

Definition 2: When $r_0 > \frac{1}{\sigma}$, the message propagation will be rapid and efficient; in contrast, when $r_0 < \frac{1}{\sigma}$, the message propagation will be slow and in control.

So if $o_0 = r_0 \frac{C}{U}$=$r_0\ \sigma$ (18)

Then $o_0 = 1$, which can serve as the critical values in RMO model to determine whether messages will be rapidly propagated.

Corollary 1: $o_0 = 1$ is the critical value in RMO mode. When $o_0 > 1$, message propagation will be rapid; when $o_0 < 1$, by contrast, message propagation will be limited.

Proof: Based on Theorem 3 and Definition 2, when $r_0 > \frac{1}{\sigma}$, i.e. $r_0 \sigma > 1$, messages will be propagated rapidly. As the Equation (18) illustrates, when $r_0\ \sigma = o_0$ >1, message propagation will be rapid. But when $o_0 < 1$, messages would not spread quickly, and $m(t)$ will decrease to approximately 0 within the time unit. Therefore, to speed up message propagation, besides increasing cluster parameter *C*, we can try to add some new services to social networks to increase user loyalty and reduce customer attrition. Moreover, some services like fan pages can be devised to directly transform receivers into messengers In other words, receivers become messengers without going through the process of oral advocacy.

Corollary 2: Assume the percent of the number of users transforming into messengers due to the participation into fan pages out of the total users on social network services to be z. When o_0 gets larger, z will be higher and message propagation will be inefficient.

Proof: Assume the percent of the number of users transforming into messengers due to the participation into fan pages out of the total users on social network service to be z. Then, r_0 can be changed into $(1-z)r_0$ and o_0 can be changed into $\overline{o_0} = \frac{C}{U}(1-z)r_0$. If $\overline{o_0} < 1$, message propagation will be limited. When $\frac{C}{U}(1-z)r_0 < 1$,

z > $1 - \frac{U}{Cr_0} = 1 - \frac{1}{o_0}$ (19)

According to Equation (19), if o_0, the percent of the members in the fan pages, gets larger, z becomes higher to contribute to rapid message propagation.

Based on the above two corollaries, the estimation of o_0 is very vital. According to Equation (18), the estimation of *C* must be conducted before the estimation of o_0, because *C* values will vary with different social network features, with different messenger conditions, with different incentive policies of social communities, and various factors in real-life situations.

Here is a method to estimate the approximate for o_0.

According to Theorem 3, when $t \to +\infty$, $m(t) \to 0$ and $r(t) \to r_\infty$. Let $m_0 + r_0 = k$, $0 \le k \le 1$, and from Equation (17),

we have $k - r_\infty + \frac{1}{\sigma}\ln(\frac{r_\infty}{r_0})$ =0 (20)

We can get the solution of Equation (20), and there is only one root solution r_∞.

Then, the solution is,

$$\frac{U}{C} = \frac{1}{\sigma} = \frac{k - r_\infty}{\ln(r_0) - \ln(r_\infty)} \qquad (21)$$

D. RMI Model

In this model, we take another factor into account: messengers may change their status into receivers but become messengers again. These messengers may change to "idle," which can be denoted by S (the separate rate). RMI model is different from RMO model, in which the change of the rate of messengers in the SNS in the time unit can be denoted by the equation: $\frac{dm(t)}{dt}$ = $Cm(t)r(t) - Um(t)$. But in RMI model that considers the factor that the rate of messengers $m(t)$ in the SNS may decrease with the increase of the separation rate, we assume that in the SNS, more messengers are more likely to become idle and the idle rate is equal to the separate rate multiplied by the rate of messengers, i.e. SM in Figure 9.

Journal Pre-proof

Journal Pre-proof

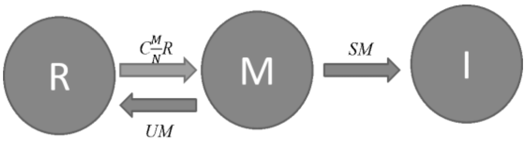


Figure 9. RMI model diagram

As a result, in RMI model, the equation is proposed as follows.

$$\frac{dm(t)}{dt} = Cm(t)r(t) - Um(t) - Sm(t)$$

However, the rate of the number of receivers in the time unit will be affected by messengers and receivers in SNS. Thus,

$$\frac{dr(t)}{dt} = -Cm(t)r(t) + Um(t)$$

So the RMI model equation can be proposed as follows,

$$\begin{cases} \frac{dm(t)}{dt} = Cm(t)r(t) - Um(t) - Sm(t) \\ \frac{dr(t)}{dt} = -Cm(t)r(t) + Um(t) \\ \frac{di(t)}{dt} = Sm(t) \end{cases} \qquad (22)$$

The message propagation curve of RMI model is similar to that of RMO model in Figure 7 and Figure 8, but outer ratio is regarded as idle ratio. The RMI model is more suitable to be applied to the social network services with poor user loyalty, which means that messengers of the fan pages may become receivers and may permanently drop out of this group.

Discussion: In the above section, we have proposed the basic RM model, and modified it into RMR, RMO and RMI models by adding into the coefficient which meets the real-life social behavior. We can also continue to extend our RMI model and add more parameters and social behavior patterns. Besides, the computation of (NP-Hard problem) will be more time-consuming. Though extending RMI model by taking more parameters and social behavior patterns into consideration may lead to more computing accuracy, it is not feasible in real-life social network applications because it will make the MPM model more complicated and require lots of computing to get the solution. Therefore, we propose the RMI model, a less complicated model, to obtain the solution more rapidly and easily and to reflect the message propagation trend on the social networks. This is exactly the objective the study wants to achieve. In this study, during the process of MPM model analyses and simulation, we found that clustering coefficient *C* plays an important key factor in the message propagation of the social networks. Thus, we will next devise a method to assess the overall clustering coefficient *C* of the MPM model to comply with the model's actual assessment and prediction accuracy in social networks, and further put the method to the future application in identifying the key social network parameters for effective message propagation.

E. The Revised MPM Model (Receiver-Light-Heavy-Messenger Model)

We can modify the original RMI model into Receiver-Light-Heavy-Messenger-Idled Model (RLHmI) model by considering the factor that some messengers might become heavy users. Therefore, the original *M* is changed into heavy users *(Hm)* and light users *(Lm)*. Hm stands for users becoming heavy users, while *Hm(t)* represents the number of users becoming heavy users within the time unit (*t*). *P* represents the percent of users transforming from light users *(Lm)* to heavy users *(Hm)*. Besides, based on some data analyses in Figure 17, it can be found that users who drop out of SNS would rarely join in once again, and that light users as well as heavy users have a chance to become idle at last. Therefore, the parameter *U* (off-line rate) in RMI model can be further divided into U_1 and $U_{2.}$ U_1 represents the percentage of users changing from light users to idle users with the time unit (*t*), while U_2 is the percent of users transforming from heavy users to idle users. S stands for separation rate in Figure 10.

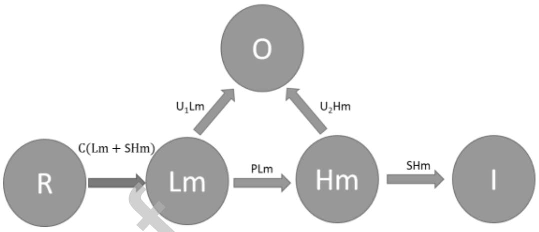


Figure 10. RMI model revised to more complex RLHmI model

Based on the assumptions given in Section III, the original variation per time unit is $\frac{dR(t)}{dt} = -C\big(Lm(t) + SHm(t)\big)$. Similarly, the original Lm variation per time unit is

$$\frac{dLm(t)}{dt} = C\big(Lm(t) + SHm(t)\big) - (P + U_1)Lm(t) = [C - (P + U_1)]Lm$$

$$\begin{cases} \frac{dLm(t)}{dt} = [C - (P + U_1)]Lm(t) + SCHm(t) \\ \frac{dHm(t)}{dt} = PLm(t) - Hm(t)(U_2 + S) \\ \frac{dR(t)}{dt} = -C\,(Lm(t) + SHm(t)) \\ \frac{dO(t)}{dt} = U_1Lm(t) + U_2Hm(t) \\ \frac{dI(t)}{dt} = SHm(t) \\ R(t) + Lm(t) + Hm(t) + O(t) + I(t) = N \\ R_0 > 0; Lm_0 > 0; Hm_0 \geq 0; O_0 \geq 0; I_0 \geq 0 \\ R(0) \approx N(0); Lm(0) \approx 0; Hm(0) = O(0) = 0 \end{cases} \qquad (23)$$

Equation (23) is a complex simultaneous differential equation. Due to its five parameters, it is difficult to get a resolution. Therefore, we could only use approximate methods to gradually get its approximate values.

Therefore, we simplify the model and just analyze the model.

$$\begin{cases} \frac{dLm(t)}{dt} = [C - (P + U_1)]Lm(t) + SCHm(t) \\ \frac{dHm(t)}{dt} = PLm(t) - Hm(t)(U_2 + S) \end{cases} \qquad (24)$$

This would be the linear system of time-invariant.

$$Hm(t) = RP = \frac{P}{\lambda_1 - \lambda_2}e^{\lambda_1 t} - \frac{P}{\lambda_2 - \lambda_1}e^{\lambda_2 t}$$

The above equation is a complex simultaneous differential equation. Due to its five parameters, it is difficult to get a resolution. Therefore, we could only use approximate methods to gradually get its approximate values. The equation is recommended to be transformed into a matrix to be analyzed with Cayley-Hamilton theorem.

Journal Pre-proof

Journal Pre-proof

IV. Simulation Environment and Result Analysis

A. *MPM Model Experiments and Simulations*

The purpose of this experiment is to make use of the RMI model of message propagation and clustering evaluation formulas proposed in this study to analyze and assess the clustering results. We hope that our analyses and evaluations can help users understand some social clustering results easily and clearly. In this experiment, the experimental and simulation platform is the IBM-compatible PC, with Intel(R) Core(TM) i7 CPU 870 @ 2.93GHz, 8GB RAM; the operating system is Windows 7; the tool-developing software is k-means with matlab 8.1 and lightspeed toolbox with epidemic spread model.

B. *Performance Comparison of Five Kinds of prediction Models (LR, RM, RMR, RMO, and RMI)*

The experiments of the proposed MPM model (i.e., RM, RMR, RMO and RMI) are conducted to evaluate the performance of different prediction models and transmission of Facebook messages. Our experimental data are collected from the research of message propagation on social networks, collected by the University of California [20]. In the experiments, the selected social network service, Facebook, has the world's largest number of users. The data are intercepted from three fan page sites during two months, ranging from October 1, 2009 to December 31, 2009. We extracted 200 Facebook multimedia topics (i.e., num_of_facebook = 200). Each topic has more than 12040 times of responses and the maximum number of responses is 22100. Besides, the message continued for at least 4 days.

In Table 1, *d* represents the average network degree, *L* indicates the average path length, *ρ* is the density of the network, and *C* represents the average social cluster coefficient. The data is obtained through actual experimental results. *U* is the average cycle of the message propagation, while *S* represents the separate rate. We divide our experiments into small-scale experiments and large-scale experiments.

Table 1. The statistical parameters of the 3 datasets

	Nodes	Edges	Action Stats	d	L	ρ	C	U	S
Data set A:	2,035	4,625	2,148	12.342	4.214	0.027	0.354	0.28	0.82
Data set B:	7,451	213,481	182,355	76.421	3.442	0.016	0.601	0.53	0.53
Data set C:	9,486	493,225	620,351	80.342	2.635	0.085	0.814	0.76	0.32

Journal Pre-proof

RM model represents the social network service without any user interaction, such as the *C* values of Email. (*C*=0.354). RMR model is more applicable to simple social network services, like blogs and social forums (*C*=0.601). RMO and RMI models are more applicable to the social network services, like Facebook, which provides users with individual fan pages (*C*=0.814). In small-scale experiments, the original Facebook data that meet the e-mail characteristics are selected as dataset A, because there are smaller *C* values in the 3 datasets. Those which meet the blog features are categorized as dataset B and Facebook features as dataset C. Through the experiments, we can confirm the scenarios in which our 4 models are suitable for various datasets. 3 Datasets are used to represent email, blog, Facebook, respectively to validate the error rates of 4 models

For Facebook_messages (*fm*), we predict its response number at every time unit starting from (training_window_size + 1) time unit to *n* time units, and define the overall prediction error rate as:

$$Predict_{error}(fm) = \frac{\sum_{t=w+1}^{n}\left|\frac{R_t-P_t}{R_t}\right|}{n-w} \tag{25}$$

In Equation (25), R_t and P_t are the real and the predictive values of *t* at a certain time unit; *w* is the training_window_size; *n* is the total number of time units, which is 200 for our simulation setting. The average prediction error rates equation is defined as in Equation (26):

$$Average\_Predict_{error} = \frac{\sum_{i=1}^{num\_of\_facebook\_messages} Predict_{error}(fm_i)}{num\_of\_facebook\_messages} \tag{26}$$

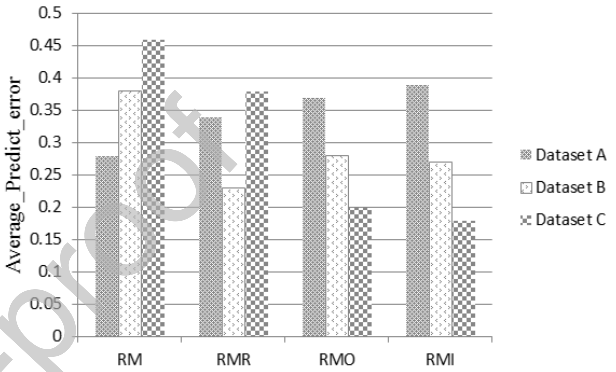


Figure 11. 3 datasets are used to represent email, blog, Facebook respectively to validate the error rates of 4 models

When dataset A, which meets the email characteristics, is used to assess the error rates of 4 models, we found that RM model has the lowest error rates of the 4 models, which also demonstrates that RM model is more suitable to describe the email scenario. In dataset B which corresponds to blog features, we found that RMR model has lower error rates, and thus RMR model is more appropriate to be utilized to describe the blog scenario. Dataset C is more in line with the Facebook features. As Figure 11 shows, RMO and RMI have the lowest error rates, while RMI model is a little bit more accurate than RMO model, since more parameters are added in RMI model to describe the actual context of Facebook. The average prediction error rates can be calculated with Equation (26).

In large-scale Facebook datasets, the larger the data scale becomes, the higher the error rate of models would be. While the average error rate in small-scale dataset is 0.18, the average error rate in large-scale dataset is 0.28 at best. The difference might result from the fact that more implicit parameters in large-scale data are ignored. Therefore, we recommend to conduct more follow-up data analyses to compute the parameters that are more suitable for the actual social network environment to enhance the accuracy of models. In the following, we use a continuous training method to predict the number of Facebook message responses at the time unit. Then, we utilize the number of Facebook message responses during the existing time units to evaluate five prediction models (i.e., LR, RM, RMR, RMO, and RMI). Linear Regression (LR) is a method for modeling the linear relationships between a scalar dependent variable and one or more explanatory variables. For our message propagation prediction, the linear prediction

Journal Pre-proof

function has a form as: *M(t + 1) = C*R(t) + U*M(t) + S*I(t)*, where parameters *C, U, S* are trained and obtained by fitting observed data in the training set. In other words, we adjust dynamically various parameters of the prediction models with the *W* time units and slide the training window size. In our experiments, the training window varies from 10 to 60 time units (i.e., training_window_size = 10, 20, 30, 40, 50, 60). The interval between two consecutive time points is set to be 40 minutes (i.e., temporal_ distance = 40 min).

Figure 12 reveals the average error rates of different prediction methods of the message propagation and the results shows that RMO and RMI have lower prediction error ratio than other three commonly used models. The complex RMI model incorporates more parameters so that it can fit in the real social network message propagation context, and the performance of the average prediction error of RMI model therefore would be the best of the five models. We also make a comparison of the time units required for the five models to conduct the complete computing of a dataset in the same environment. The result indicates that the RMI model – incorporating more parameters and requiring a considerable amount of matrix calculation- produces the lowest error rate but takes comparatively more time units compared with the other four models, while LR model, the simplest model, requires comparatively less time. The time units that the other 4 models take fall in-between. The RMI model, compared with the other three models, has a smaller average error prediction rate; besides, the time units it takes are also within the acceptable range. Considering the required computing time and the average error prediction rate, RMI model is recommended for use in the social networks recommendation system which requires instant feedback.

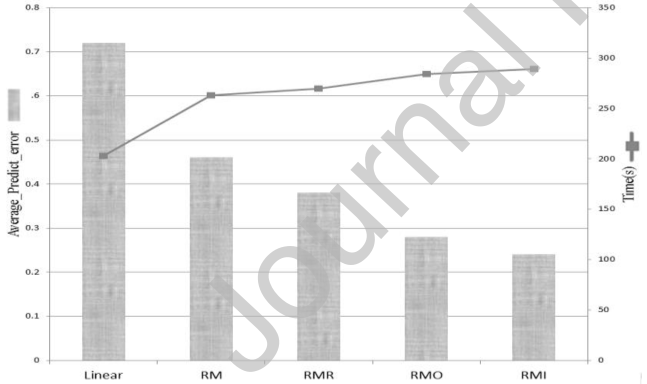


Figure 12. $Average\_Predict_{error}$ and time units of different prediction methods (training_window_size =40, temporal_distance = 40 mins)

C. Effects of Training Window Size with Four Prediction Methods (RM, RMR, RMO, and RMI)

In the experiment, we use the large-scale datasets and further compare the RMO and RMI performances of different training window sizes, ranging from time units 10 to 60. In Figure 13, we find that the more training time units, the more Facebook feedback recommendation information can be obtained. RMO and RMI performances are always better than those of other models. At time unit 10, 20 and 30, RMO has the best performance than the other models because RMO is more robust. After the time unit 40, the performance of RMI model would be better than those of other models, since RMI model has had more training of datasets. The performances of all models decline after time unit 40 because the duration of a Facebook message is not very long, approximately 3-5 days. At the later stage, the number of Facebook messages reduces, so even a very small error of prediction value may incur a bigger prediction error. It can demonstrate that prediction models will have a certain degree of accuracy in the early stage of prediction, but over time, the prediction error rate will grow.

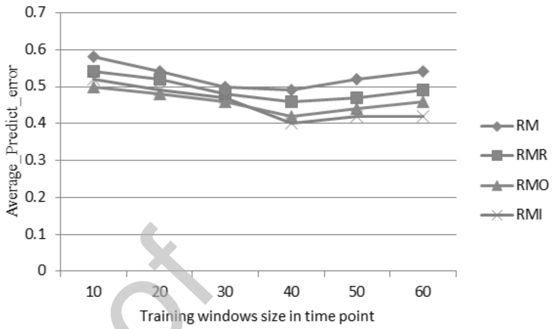


Figure 13. Prediction results of RM,RMR, RMO and RMI vs. training window sizes (temporal_interval = 40 mins)

D. Comparison of Performances of RMO and RMI

In this comparison, we use the large-scale dataset and examine the effects of parameter *C* setting in Equation (5). In Figure 14, the curve with solid line denotes the real Facebook message numbers, the curve with diamond dashed line represents the basic setting in RMO, and the curve with circle double dashed line denotes the advanced setting in RMI. As shown in the figure, around the first peak, at the first three time units, RMI has a better prediction performance than RMO, because the *S* setting of RMI method increases the transmission rate of RMI, which is faster than RMO. However, around the 44th time unit, RMO performs better because when real Facebook message numbers are small, they are more sensitive to some random factors in RMI. A similar prediction performance can be seen at the 45th to 48th time units. Overall speaking, RMI with more parameters can make a better prediction than RMO with *C* setting when Facebook message numbers are large. However, when real Facebook message numbers are small, RMO is better than RMI. This result illustrates the fact that Facebook message propagation on social networks preserves the characteristics similar to those of the spread of traditional infectious disease.

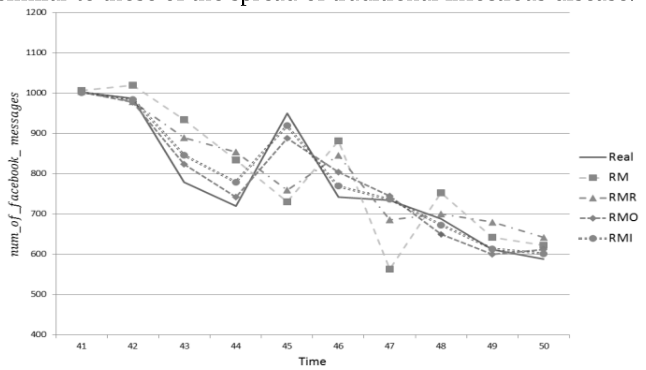


Figure 14. Prediction results of RM, RMR, RMO and RMI

Journal Pre-proof

Journal Pre-proof

compared with real-life conditions (training_window_size = 40, temporal_interval = 40 mins)

E. Effects of Training Window Size with RMO and RMI

As illustrated in Figure 15, prediction error rates of both RMO and RMI decrease with the rise of the training window sizes (i.e., number of time units). This result is consistent to those results obtained by using the other prediction models in Figure 13. The more training time units, the more Facebook message feedback information can be gained, and the influence of some noisy time units can be eliminated to obtain more appropriate Facebook message trends. The figure shows that RMI performs better than RMO when the training window size is more than 40 time units because RMI model adds other parameters to its evaluation method, which would better grasp the inherent characteristics of the Facebook message transmission rate.

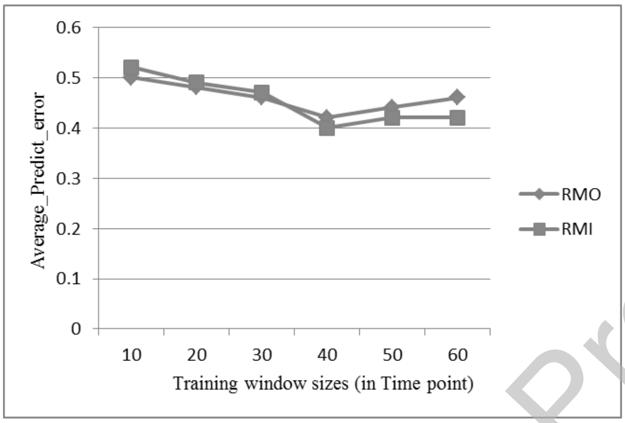


Figure 15. Average_Predict_error of RMO and RMI vs. training window sizes (temporal_interval = 40 mins)

In this experiment, the phenomenon of over-fitting can be noticed. It might result from the use of too many parameters while we make adjustments to the MPM model. During the process of over-fitting, when the amount of training data increases, the performance in the application to the unknown data gets worse. In order to avoid over-fitting, we can utilize some additional methods such as cross-validation data analysis for early studying, to pinpoint the time when further training do not lead to better results. But another phenomenon results from the use of too few parameters, which is called under-fitting. Under-fitting occurs when a model does not fit the data well. To tackle these problems caused by overfitting and underfitting, we utilize the big data analyses to change the previous method of “looking for data based on problems” into the current one of “looking for problems based on data” to explore the available values, find the appropriate parameters and make proper adjustments to models.

F. Effects of Training Window Sizes with the Time Interval Between Two Consecutive Time Units

As Figure 16 shows, the average prediction error rates of RMO and RMI models decrease when the temporal interval between every two time units increases from 10 minutes to 60 minutes. This result is consistent to the one shown in Figure 13 when the train window size gets enlarged; the effect is similar to enlarging the temporal scope of training window size. Also, RMI performs better than RMO when the temporal interval is over 40 minutes.

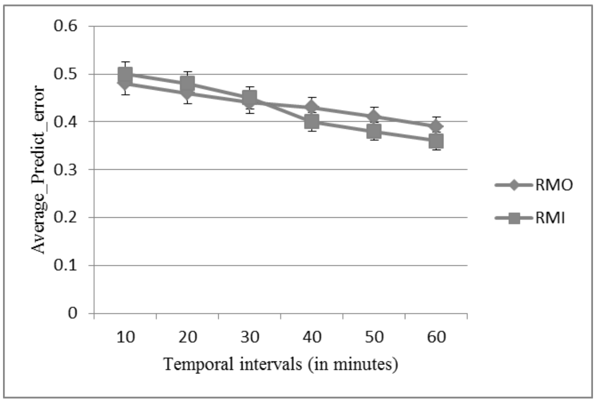


Figure16. Average_Predict_error of RMO and RMI vs. temporal intervals
(training_ window_size = 40)

In this paper, we propose MPM message propagation models developed on the social networks to analyze and predict social message propagation in Facebook. We extend and modify the basic RM message propagation model for analyzing Facebook message propagation. Two different kinds of parameter setting methods are utilized in examining the performances of different models. Our performance results illustrate that the extended RMI model is more effective in predicting social message propagation than the other three commonly used prediction methods. We find two similar fan pages on Facebook: one is with more heavy users, while the other is with most light users- to make a comparison of message propagation in Figure 17.

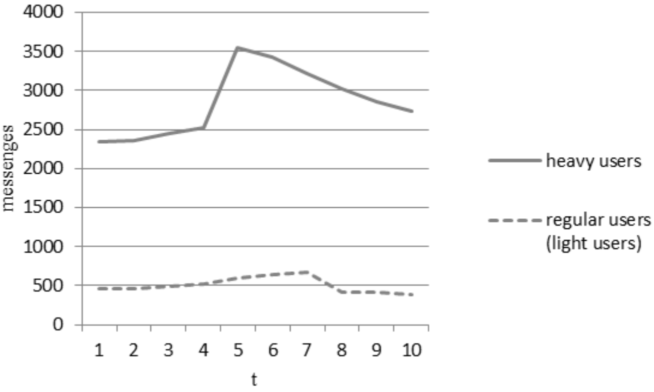


Figure 17. Comparison of message numbers propagated by heavy users and light users

Based on the data analysis, in the social communities with large numbers of heavy users, the message propagation would be more rapid, reaching a peak at some time; besides, the message numbers would decrease at a slower speed. In contrast, the social communities with most light users would have slower message propagation, an unobvious peak, and a rapid message number decrease. In conclusion, heavy users would share more of their feelings and emotions with other Facebook users than light users. In other words, heavy users have stronger influence on Facebook; they would make interaction through posts, graffiti wall posts and chat rooms. Judging from this, heavy Facebook users would cause the increase of message numbers

Journal Pre-proof

and enhance the message propagation speed. With the result, we would make some adaptations to our MPM model to make it more applicable to the real-life environments. If MPM model can be applied and modified according to data analyses, its accuracy would be enhanced and would be closer to real-life scenario (the solid real line and the circle double line) in Figure 18 and Figure 19.

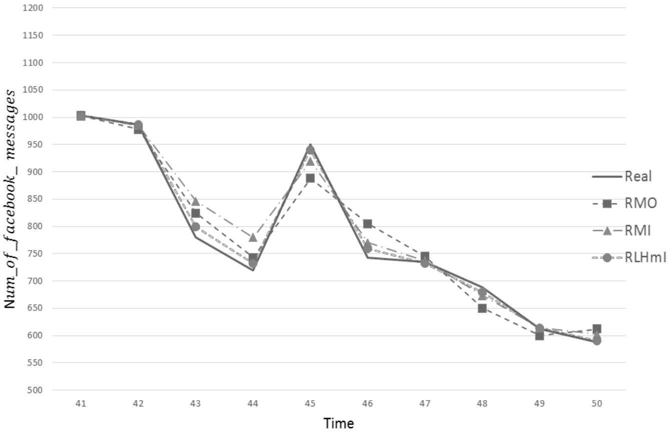


Figure 18. Prediction results of RMO ,RMI and revised RLHmI model compared with the real-life condition (training_window_size = 40, temporal_interval = 40 mins)

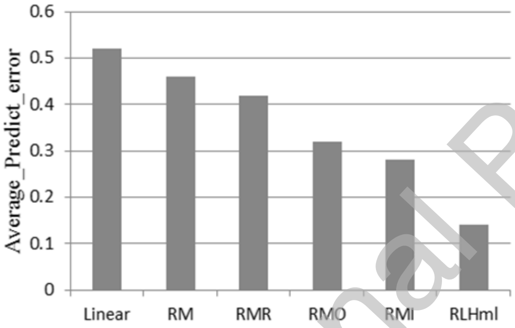


Figure 19. Average_Prediction_error and Time of different prediction methods with RLHMl model

Based on the above discussion, models can be analyzed through big data to find more parameters to revise our MPM model, making them more applicable to real-life conditions. But too many parameters may bring about the problem: how can we find the resolution of the complex model? It is also an NP hard problem. Therefore, in the future practical application, it is recommended to find a model whose condition is close to the social network services. Besides, some practical application can be devised, such as the real time social recommendation system, which can consider the rapid response time. With this setting, finding the message propagation trend and being able to respond to users' demands are more important than a complex model. MPM model is one such applicable example. If MPM model can be applied and modified according to data analyses, its accuracy would be enhanced and would be closer to real-life scenario. In the large-scale Facebook dataset experiments, the modified RLHml model can enhance forecast accuracy. Nowadays, users are spending more time making interaction on social communities, which also contribute to the rapid accumulation of data. The hundreds of millions of pieces of data have posed a severe challenge to information retrieval, processing and analysis. Due to the rapid growth and great variability of data, the big social data usually cannot be accessed at one time. Besides, because the data size is calculated in Giga or even Tera, they cannot be read into memory for further processing. Therefore, it is of vital importance to devise a systematic, effective, and efficient method to retrieve the data which can be representative of all data to do the follow-up processing. One method to save and analyze the big data on the social network is "simplification." The purpose of simplification is to make the simplified social network similar to the proposed model and yet still representative of the big data. That is, the simplified data can still retain the statistical characteristics, and can produce as few errors as possible in the simplification process.

V. Conclusion

Because the social network is a complex system, there are many variables in the change in interlocking layers. Take the Butterfly Effect, a chaotic way and fraction in the social network as an example. Everything looks disordered, but there still exist some kinds of order or trends in between. Traditional mathematics, such as differential equations, is a complex phenomenon which cannot be calculated. We hope to find new ways to analyze these phenomena, especially the chaotic behavior, which is a very important topic in nonlinear science due to its potential applications in social networks. In this research, we have reviewed some recent works on the structure and function of social network service systems. Research in this area has been motivated to a high degree by empirical studies of real and virtual world social networks such as the Internet, the World Wide Web, social networks, collaboration networks, citation networks, and a variety of biological networks. We have reviewed these empirical studies in Section II, focusing on a number of model characteristics of social networks. The largest portion of this research has been focused on discussions of these message propagation models in Section III. We also described the mathematic models of the social network services. We proposed MPM model, which include RM model, RMR model, RMO model and RMI model, to explain the message propagation behaviors on the social networks. We can make use of the proposed models to evaluate and analyze real-life conditions on social networks, including the behavior of message receivers and messengers. Finally, the MPM model can be applied and modified according to Hadoop data analysis, and its accuracy would be enhanced and would be closer to real-life scenarios. In the large-scale Facebook dataset experiments, the modified RLHml model can enhance forecast accuracy. Inspired by these observations, many researchers have proposed models of social networks that typically seek to explain either how networks come to have the observed structure, or what the expected effects of that structure will have. However, there is still much to be done in developing more sophisticated models of networks, both to help us understand network topology and to act as a substrate for the study of processes taking place on networks. While some network characteristics, such as degree distributions, have been thoroughly modeled and their causes and effects well understood, others such as correlations, cluster coefficient, and

Journal Pre-proof

Journal Pre-proof

community structure have not. It seems that these characteristics will affect the behavior of networked systems substantially, so our current lack of suitable techniques to handle them leaves a large gap in our understanding. If we can gain such understanding, it will give us new insight into a vast array of complex and previously poorly understood phenomena.

But as far as mobile social network services are concerned like what's app, wechat, line and twitter [21], the parameters in the various analysis perspectives still require further conceptualization and definition in order to more accurately capture the characteristics of social network services and thus provide a higher reference value. Since the study primarily works on the compilation of the analysis concepts of Facebook, more focuses would be directed on Facebook-oriented research in the process of constructing the analysis framework. Besides, due to the limited research time and data, the observation parameters in each analysis require further discussion and fail to testify the concept and its application scope. Future related works are recommended to focus on the parameters of the study's analysis framework, clarifying the definitions and contents of the concept, and further expanding or refining applicable parameters.

Moreover, the study only utilizes the currently popular social network service [22], Facebook, to testify the applicability of the analysis framework without considering the large diversity of current social network services. However, the applicability of the framework should be testified through more social network services to confirm its application value. Therefore, future research could attempt to apply this framework to more mobile social network services, with a view to identify the framework's disadvantages and further modify and refine the framework. It is recommended that Hadoop can possibly be replaced with Spark [23], a more real-time and rapid analysis platform. Spark is a more flexible operational framework, suitable for various types of applications such as batch processing, workflow, interaction analyses, and traffic flow processing. Therefore, in the future, Spark can replace Map/Reduce, since Spark can serve as a widely applied computing engine, and it is also more suitable for use in real-time mobile social network services. The future works can also make use of the data visualization tools to help analyze the results, identify the associations between variables, and even predict future trends to contribute to more analysis efficiency.

Author statement

- Dr. Wei Kuang Lai is responsible for the innovation of the paper structure design.

 Dr. Yi Uan Chen is responsible design our method from the five models in this article.

 Dr. Tin-Yu Wu is responsible for analyzing and compare the five models to show the results.

Conflict of Interest

None.

REFERENCES

[1] Sam G.B. Roberts, Ruth Wilsonc, Pawel Fedureka, R.I.M. Dunbarb, "Individual differences and personal social network size and structure", Personality and Individual Differences, Vol.44, pp.954-964, 2008.

[2] Elihu Katz and Paul Felix Lazarsfeld, "Personal influence: The part played by people in the flow of mass communications". Piscataway: Transaction Publishers, 2006.

[3] Pedro Domingos and Matt Richardson,"Mining the network value of customers".Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: pp. 57-66, 2001.

[4] Matthew Richardson and Pedro Domingos, "Mining knowledge-sharing sites for viral marketing".Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada: pp. 61-70, 2002.

[5] David Kempe, Jon Kleinberg, Éva Tardos, "Maximizing the spread of influence through a social network".Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D.C, USA: pp. 137-146, 2003.

[6] David Kempe, Jon Kleinberg, and Éva Tardos, "Influential nodes in a diffusion model for social networks".32nd International Colloquium on Automata, Languages and Programming. Lisbon, Portugal: pp.1127-1138, 2005.

[7] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan, "Learning influence probabilities in social networks".Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA: pp.241-250, 2010.

[8] Chi Wang, Jie Tang, Jimeng Sun, and Jiawei Han, "Dynamic social influence analysis through time-dependent factor graphs".Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. Kaohsiung City, Taiwan: pp. 239-246, 2011.

[9] David Kempe, Jon Kleinberg, and Éva Tardos, "Influential nodes in a diffusion model for social networks".32nd International Colloquium on Automata, Languages and Programming. Lisbon, Portugal: pp.1127-1138, 2005.

[10] Adrien Guille, Hakim Hacid, Cecile Favre, Djamel A. Zighed, "Information diffusion in online social networks: A survey". SIGMOD Record,42 (2): pp.17, 2013.

[11] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst, "Cascading behavior in large blog graphs".Society for Industrial and Applied Mathematics International Conference on Data Mining. Minneapolis, USA, 2007.

[12] Jacob Goldenberg, Barak Libai, and Eitan Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth". Marketing Letters, 12 (3): pp.211-223, 2001.

Journal Pre-proof

Journal Pre-proof

[13] Daniel Gruhl, R. Guha, David Liben-Nowell, Andrew Tomkins, “Information diffusion through blogspace”.Proceedings of the 13th International Conference on World Wide Web. New York, USA: pp. 491-501, 2004.
[14] David Kempe, Jon Kleinberg, Éva Tardos, “Maximizing the spread of influence through a social network”.Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D.C, USA: pp. 137-146, 2003.
[15] Charles H. Hubbell, “An input-output approach to clique identification”. Sociometry, 28: pp.377-399, 1965.
[16] Mark Granovetter, “Threshold models of collective behavior”. American Journal of Sociology, 83 (6): pp. 1420-1443, 1978.
[17] Wei Chen, Yifei Yuan, and Li Zhang, “Scalable influence maximization in social networks under the linear threshold model”.Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney, Australia: pp.88-97, 2010.
[18] David Kempe, Jon Kleinberg, and Éva Tardos, “Influential nodes in a diffusion model for social networks”.32nd International Colloquium on Automata, Languages and Programming. Lisbon, Portugal: pp.1127-1138, 2005.
[19] Frank M. Bass, “A new product growth model for consumer durables”. Management Science, 15 (5): pp.215-227, 1969.
[20] Minas Gjoka, Carter T. Butts, Maciej Kurant, and Athina Markopoulou,” Multigraph Sampling of Online Social Networks” IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 29, NO. 9, 2011.
[21] William H. Woodall, Meng J. Zhao, Kamran Paynabar, Ross Sparks and James D. Wilson. “An overview and perspective on social network monitoring” IIE Transactions: pp.20-30, 2016.
[22] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ Global social networks ranked by number of users 2016
[23] Shivaram Venkataraman, Zongheng Yang, Davies Liu, Eric Liang, Hossein Falaki, Xiangrui Meng, Reynold Xin, Ali Ghodsi, Michael Franklin, Ion Stoica, and Matei Zaharia. “SparkR: Scaling R Programs with Spark”, SIGMOD 2016.
[24] Tin-Yu Wu, Nadra Guizani, Jhih-Siang Huang, "Live Migration Improvements by Related Dirty Memory Prediction in Cloud Computing", Journal of Network and Computer Applications (JNCA), Vol.90, pp. 83-89 July 2017.
[25]Tin-Yu Wu, Wen-Kai Liu, "Game Theory-based Global Optimization for Inter-WBAN Interference Mitigation", Wireless Communications and Mobile Computing (WCMC), Vol. 16, Issue 18, pp. 3439-3448, December 2016.

W. K. Lai received a BS degree in Electrical Engineering from National Taiwan University in 1984 and a Ph. D. degree in Electrical Engineering from Purdue University in 1992. He joined the faculty of Department of Computer Science and Engineering, National Sun Yat-Sen University in August 1992 and is now a professor. His research interests are in high-speed networks and wireless networks.

Yi-Uan Chen received the M.S. degree in electrical engineering from National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C., in 2001. He received his Ph.D. degrees in the the Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan, R.O.C. He joined Chunghwa Telecom Co., Ltd. Taipei, Taiwan, R.O.C., in August 1999, and is now a lecturer of telecom training center. His research interests are in wireless networks, social network analysis and cloud computing.

Tin-Yu Wu currently works as an Full Professor in the Department of Computer Science & Information Engineering, National Ilan University, Taiwan. He received his M.S. and Ph.D. degrees in the Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan in 2000 and 2007 respectively. His research interests focus on the big data analytics, cloud computing and mobile computing.

Journal Pre-proof