Physica A 415 (2014) 87–94

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Viral information propagation in the Digg online social network

Mark Freeman [a], James McVittie [b], Iryna Sivak [c,1], Jianhong Wu [d,*]

[a] *Program for Evolutionary Dynamics, Harvard College, One Brattle Square, Suite 6, Cambridge, MA 02138-3758, USA*
[b] *Department of Statistics, University of Toronto, 100 St. George St., Toronto, Ontario, Canada*
[c] *Department of Mathematics, Taras Shevchenko National University of Kyiv, Volodymyrska st. 64, 01601, Kyiv, Ukraine*
[d] *Laboratory for Industrial and Applied Mathematics, York University, 4700 Keele Street, Toronto, Ontario, Canada, M3J 1P3*

HIGHLIGHTS

- We propose and analyze an epidemiological model for information propagation in online social network.
- We characterize peak timing, turning point, viral period, and final size of the number of votes.
- There are significant similarity and difference between information propagation in OSNs differs from disease spread in populations.
- Simple dynamic models can provide accurate prediction of information propagation in OSNs.

ARTICLE INFO

Article history:
Received 27 November 2013
Received in revised form 2 June 2014
Available online 16 July 2014

Keywords:
Online social network
Digg network
Information spread
Mathematical epidemiology
Richards model

ABSTRACT

We propose the use of a variant of the epidemiological SIR model to accurately describe the diffusion of online content over the online social network Digg.com. We examine the qualitative properties of our viral information propagation model, demonstrate the model's applications to social media spread in online social networks with particular focus on accurately predicting user voting behavior over a period of 50 h. The model allows us to characterize the peak time, turning point, viral period and final size (total number of votes), and gives much improved prediction of user voting behaviors than other established models.

1. Introduction

Everyday in online social networks (OSNs), thousands of users post news articles, videos, photos etc. which become visible to their connected users as new online content. As most of these forms of media never spread to a wide audience from the sources, many users are influenced by their networking. The causes and dynamics by which information proliferates throughout OSNs are still poorly understood. A greater comprehension of the mathematics underlying the spread of information in OSNs would have important applications for advertisers seeking to wage more effective online marketing campaigns and may enable a more rapid spread of information over OSNs in the aftermath of political crises or natural disasters.

The focus of this article is the OSN Digg.com (DOSN). In this network, users are able to post content to a personal web page, vote for ("digg") or against ("bury") this content and share the content with users to whom they are connected. There

[*] Corresponding author. Tel.: +1 4167365356.
E-mail address: wujh@mathstat.yorku.ca (J. Wu).

[1] Current address: Centre for Complexity Science, University of Warwick, Coventry CV4 7AL, United Kingdom.

http://dx.doi.org/10.1016/j.physa.2014.06.011
0378-4371/© 2014 Elsevier B.V. All rights reserved.

are two forms of connections between users: directed (one user can share content with another user but not vice versa) and bidirected (both users can share content with each other). Once posted content receives some large number of votes over a particular period of time, the content is then posted to the homepage of Digg.com and is visible to all users in the DOSN.

One of the reasons we select DOSN to illustrate our epidemiological approach is our accessibility to the dataset of Ref. [1] which contains the information of voting characteristics of the DOSN for June 2009. In particular, the data contains 3553 distinct stories (online content), the number of votes a particular story received, the particular users that voted for each story and the time at which each user cast the vote. On average, each story received approximately 850 votes where the minimum number of votes was 122 and the maximum 24 099. It should be noted, that this dataset only includes the stories that were promoted to the homepage of Digg.com in June 2009. In addition to voting data, [1] also contains the connectivity information of 71 367 distinct users which includes: the users to which each user is connected, the time at which the connection was created and the type of connection that was created (directed or bidirected). We determined that on average, every user is connected to 24 other users of which approximately half were directed (48.901%) and half were bidirected (51.099%). As in Ref. [2], we defined the distance metric between two users in the DOSN as the minimum number of connections (directed or bidirected) needed to connect them. We defined two users as being disconnected if there does not exist any path in the DOSN connecting them. We refer to Ref. [3] for the excellent empirical characterization of this data.

A goal of modeling the propagation of information in an OSN is to understand the rate at which a piece of online content influences the users as a function of time and distance away from the source of the propagation. The linear diffusive model of Feng et al. [4] used a temporal–spatial partial differential equation (PDE) model to explain these rates of spread in the DOSN. By fixing their model's parameters and altering the initial conditions to replicate the information propagation, they were able to achieve an average model accuracy of 97.41% for the most popular story. Additionally, they examined all stories receiving more than 3000 votes (134 stories). In approximately 60% of these stories, they had model accuracies greater than 80%.

Our focus was directed towards an adaptation and application of an epidemiological model which describes the spread of a virus in a population. In modifying and utilizing this model, we were able to predict the cumulative number of users who voted for any shared story at time t (hours) after its initial posting, the time period (viral period) during which the story diffuses quickly through the DOSN, and the peaking time for the total time of "influence users" to reach the maximum, and the turning point when the information spread starts to slow down. By using this model, we achieved higher model accuracies than [4] in both the most popular story and the most popular 134 stories. Furthermore, we achieved an average predictive accuracy of approximately 80% for all voted stories.

2. Main results

2.1. The model

We modeled the diffusion of a particular story through the DOSN by using a modified epidemiological SIR model [5]. As in modeling the spread of a virus in a population, we used similar SIR definitions from epidemiology to categorize the users of the DOSN at any given time in relation to any given story. The "susceptible" population $S(t)$ is comprised of users who have not yet voted for a particular story, the "infected" population $I(t)$ consists of users who have voted for a particular story and are visible on their connected users' homepages and the "recovered" population $R(t)$ is composed of users who have voted for a particular story and (after a period of time) are no longer visible on their followers' homepages thus having a negligible influence in sharing the story. We will focus on the cumulative number of people $C(t)$ who have ever voted for a particular story, but we will also examine the temporal evolution of the number of "infected" users.

We made the assumptions that if the spread of a story to different users through the promotion to the Digg homepage is neglected, then the rate of spread of the cumulative number of users, $C'(t)$, for a given story is equal to the product of the number of existing (directed or bidirected) network connections from infected to susceptible individuals and the per-connection hourly rate of story successful spread. Since all the high voted Digg stories from Ref. [1] eventually became promoted to the Digg homepage but were ultimately not voted for by less than 1.19% of the DOSN, we conclude that $C'(t)$ is approximately proportional to the total number of outward directed connections from the infected individuals. Because the total number of directed connections leading from infected users is proportional to the number of infected individuals, it follows that

$$C'(t) = \beta(t)I(t),$$

where $\beta(t)$ is the number of hourly connections times the per-connection successful spread rate of the story. We assumed the spread rate exhibited an exponential decay over time t, which yielded

$$C'(t) = (\beta_0 e^{\alpha t} + c_0)I(t),$$

where $\beta_0, c_0,$ are positive constants and α is a negative constant. By making the final assumption that a constant fraction σ of infected individuals "recover" (cease being potential sources of story spread) per hour, we arrived at the following system of equations:

$$S'(t) = -(\beta_0 \mathrm{e}^{\alpha t} + c_0)I(t) \tag{2.1}$$

$$I'(t) = (\beta_0 \mathrm{e}^{\alpha t} - \delta)I(t) \tag{2.2}$$

$$R'(t) = \sigma I(t), \tag{2.3}$$

where $\delta = \sigma - c_0, \sigma$ is a positive constant, and $c_0 < \sigma$. Note that the recovered user compartment $R(t)$ is decoupled from the first two equations (the *SI*-dynamics), but it is useful to explicitly introduce this compartment so that there is a specific meaning to the parameter σ.

It is easy to observe that $S(t) + I(t) + R(t) = N$, with N being the number of users in the DOSN. It is also easy to notice that $C(t) = N - S(t) = I(t) + R(t)$ at any given time. With these observations, we can show that the above model system, has the solution in a closed form:

$$S(t) = N - C(t) \tag{2.4}$$

$$I(t) = \mathrm{e}^{\frac{\beta_0}{\alpha}(\mathrm{e}^{\alpha t}-1)-\delta t} \tag{2.5}$$

$$R(t) = \frac{c_0\sigma}{\alpha}\mathrm{e}^{-\frac{\beta_0}{\alpha}}\left(-\frac{\beta_0}{\alpha}\right)^{\frac{\delta}{\alpha}}\Gamma\left(\frac{-\delta}{\alpha}, -\frac{\beta_0}{\alpha}\mathrm{e}^{\alpha t}, -\frac{\beta_0}{\alpha}\right) \tag{2.6}$$

$$C(t) = 1 + \int_0^t (\beta_0 \mathrm{e}^{\alpha x} + c_0)\mathrm{e}^{\frac{\beta_0}{\alpha}(\mathrm{e}^{\alpha x}-1)-\delta x}\mathrm{d}x, \tag{2.7}$$

this last component can be further expressed as

$$C(t) = 1 + \mathrm{e}^{-\frac{\beta_0}{\alpha}}\left(-\frac{\beta_0}{\alpha}\right)^{\frac{\delta}{\alpha}}\left[\Gamma\left(\frac{-\delta}{\alpha}+1, -\frac{\beta_0}{\alpha}\mathrm{e}^{\alpha t}, -\frac{\beta_0}{\alpha}\right) - \frac{c_0}{\alpha}\Gamma\left(\frac{-\delta}{\alpha}, -\frac{\beta_0}{\alpha}\mathrm{e}^{\alpha t}, -\frac{\beta_0}{\alpha}\right)\right],$$

where the generalized incomplete gamma function is defined as

$$\Gamma(x, a, b) = \int_a^b t^{x-1}\mathrm{e}^{-t}\mathrm{d}t.$$

Similar to the epidemiological Richards model [6], $\beta(t)$ represents the "growth rate" in the number of users that were "infected" by a particular story. It is the relationship between the Richards and the epidemiological compartmental SIR model from Ref. [6] that motivates this study using a viral spread model to predict the propagation of the story from a particular user (the source) to the users in the DOSN.

2.2. Properties of the model

2.2.1. Long term behavior and turning point

For examining the long term outcome of this "epidemiological" process of the information propagation, we denote $C_\infty = \lim_{t\to+\infty} C(t)$. We obtained the following

$$C_\infty = 1 + \mathrm{e}^{-\frac{\beta_0}{\alpha}}\left(-\frac{\beta_0}{\alpha}\right)^{-\frac{c_0-\sigma}{\alpha}}\left[\Gamma\left(\frac{c_0-\sigma}{\alpha}+1, 0, -\frac{\beta_0}{\alpha}\right) - \frac{c_0}{\alpha}\Gamma\left(\frac{c_0-\sigma}{\alpha}, 0, -\frac{\beta_0}{\alpha}\right)\right].$$

Consequently, the number of infected individuals goes to zero as time increases to infinity because of the σ recovery constant and its larger size relative to that of the c_0 infection constant in the $\beta(t)$ growth rate expression. As in Ref. [6], we defined the "turning point" of a story as the time at which the rate of spread ceases to increase and begins to decrease. This represents the critical point in time when a particular story's influence has changed from being strong to being weak because the rate of infection of this story has begun to decrease. This turning point only occurs if $C''(t) = 0$ and occurs at time:

$$t_{\mathrm{turn}} = \frac{1}{\alpha}\ln\left(\frac{-\alpha + \sigma - 2c_0 + \sqrt{(\alpha + 2c_0 - \sigma)^2 - 4c_0(c_0 - \sigma)}}{2\beta_0}\right)$$

which occurs only if

$$\beta_0 > \frac{-\alpha + \sigma - 2c_0 + \sqrt{(\alpha + 2c_0 - \sigma)^2 - 4c_0(c_0 - \sigma)}}{2}.$$

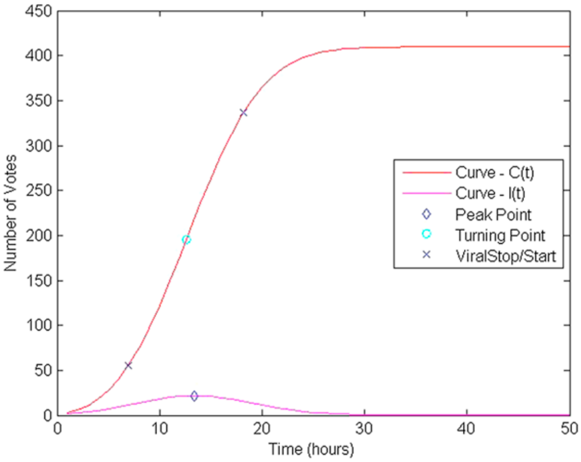


Fig. 1. Visualization of key "epidemiological" timings for a particular story: peak infection time, turning point, the starting and ending times of viral period.

2.2.2. Virality of a story

We define the "viral period" of a story as the period of time during which the particular story spreads rapidly across the social network. This period begins and ends when $C'''(t) = 0$. As shown below, this cubic equation $C'''(t) = 0$ has at most two solutions that are real numbers. If there are two positive real solutions, this implies that there is a viral starting time and viral stopping time. If there is one real solution, this implies the existence of a viral stopping time where by extrapolation, the viral starting time would be negative. Finally, if there is no positive solution, then there is no real "outbreak" of the spread of the story.

We now show that $C'''(t) = 0$ has at most two positive real solutions. Taking the derivative of the equation $C'(t) = 0$ twice, we end up with the following cubic equation

$$\phi(B) = B^3 + (3\alpha + 3c_0 - 2\sigma)B^2 + (\alpha^2 + 3\alpha c_0 + 3c_0^3 - 2\alpha\sigma - 4c_0\sigma + \sigma^2)B + (c_0^3 - 2c_0^2\sigma + c_0\sigma^2) = 0,$$

where $B = \beta_0 \mathrm{e}^{\alpha t}$. Notice that the leading coefficient and constant terms are both positive, then $\lim_{B\to -\infty} \phi(B) = -\infty$ and $\phi(0) > 0$. Then an application of the intermediate value theorem to the continuous function of ϕ ensures that $\phi(B) = 0$ must have at least one negative zero. This implies the existence of at most two real value solutions.

2.2.3. Peak infection

We define the "peak infection" time (peaking time) of a story as the point in time when $I(t)$ reaches its maximum. This is equivalent to the point in time when a story can infect its maximum number of users in the DOSN. This value is calculated by solving $0 = I'(t)$, where the peak infection time, denoted by t_{peak}, is given by

$$t_{\text{peak}} = \frac{1}{\alpha} \ln\left(\frac{\sigma - c_0}{\beta_0}\right)$$

and this occurs only if

$$\sigma - c_0 < \beta_0.$$

This restriction ensures that the peaking time is always positive if it exists. If the restriction is not satisfied, then there would never be a peak infection time and the number of infected users always decreases from $t = 0$. (In infectious disease epidemiology, this corresponds to the case where the basic reproduction number is less than the unity, and hence there is no disease outbreak even if an infectious individual is initially introduced into the population, see Ref. [5].)

Fig. 1 demonstrates various critical times to characterize the story: the starting and ending times of the viral period, the peaking time and the turning point. A number of stories in Ref. [1] exhibit the existence of these critical times; Fig. 1 is based on story #42.

2.3. Methods and results

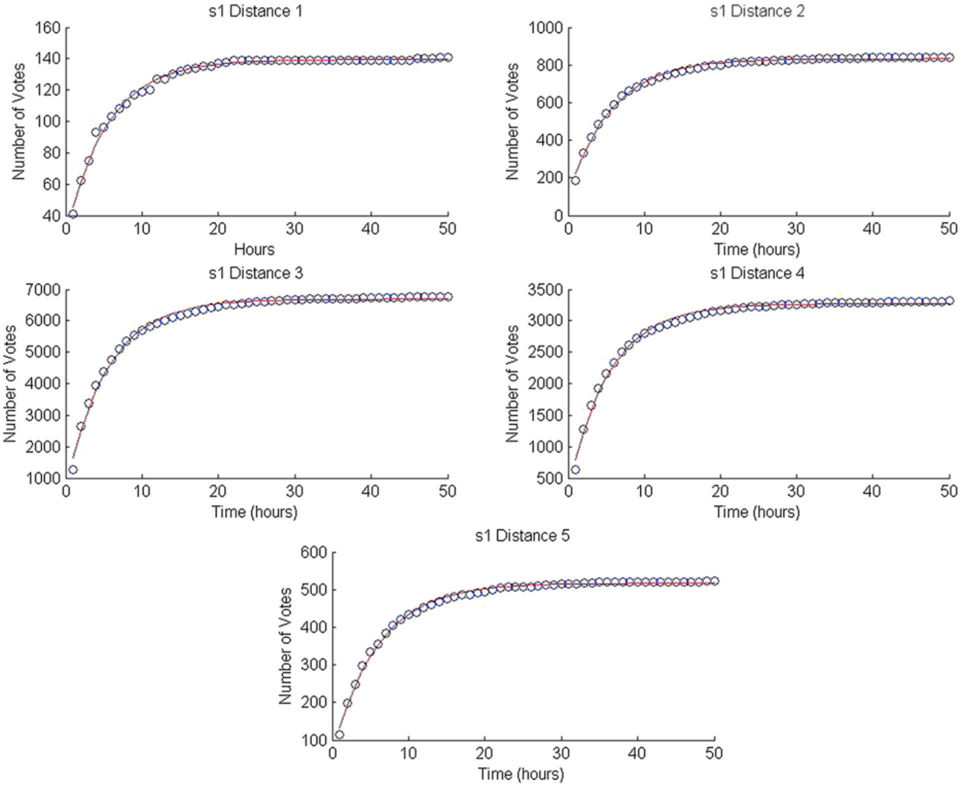
2.3.1. Data formatting

We analyzed the above model $C(t)$ (integral version) by fitting it to Ref. [1]. As in Refs. [4,7], the data was formatted over a period of 50 h. At the end of each hour, we obtained the cumulative number of users who voted for the story and used that value as our measurement of voting density. As in Ref. [2], we used the following measure to determine the model accuracy:

$$1 - \frac{|\text{predicted value} - \text{actual value}|}{\text{actual value}},$$

Table 1
Predictive accuracies for s1 over distance.

Distance	Average	Time 1	Time 2	Time 3	Time 4	Time 5
1	99.03%	97.07%	99.09%	99.53%	99.84%	99.53%
2	98.57%	96.26%	98.56%	99.41%	99.56%	99.18%
3	98.22%	94.92%	98.40%	99.52%	99.64%	99.04%
4	98.28%	95.15%	98.36%	99.45%	99.56%	99.04%
5	98.78%	96.87%	98.73%	99.47%	99.61%	99.31%
Overall	98.576%	96.054%	98.628%	99.476%	99.642%	99.22%


s1 Distance 1

s1 Distance 2

s1 Distance 3

s1 Distance 4

s1 Distance 5

Fig. 2. Fits of curve of cumulative votes to story s1 over the friendship hop distances 1–5.

where the least squares optimization, with a maximum number of 100 000 numerical iterations, was used to determine the four optimal parameter estimates σ, c_0, α and β_0 to minimize |predicted value − actual value|.

2.3.2. A case study of the most popular story

We performed a case analysis of the most popular news article, denoted by s1, to determine whether $C(t)$ is appropriate for high voting density stories. The cumulative number of votes plotted against time for s1 exhibited a logistic relationship. In fitting our particular model, we achieved higher levels of predictive accuracy than the previous linear diffusive model in Ref. [4]. In Ref. [4], the friendship hop distance metric was used, this is the natural metric of distance between two users defined as the length of the shortest path, measured by the number of hops from one user to another in the social network graph. On average, our predictive accuracy was over 98% over all distances away from the source (see Table 1). Moreover, for all distances, we achieved a nearly perfect fit (>99%) for the period of 20–50 h after the initial vote was cast for s1.

Unlike the previous models [2,4] which held their respective parameters fixed and altered the initial conditions depending on the data, the parameters of this model were optimized independently over each distance from the source. We plotted, in Fig. 2, the number of votes for s1 over all 5 distances and plotted this curve along with the points.

Our model does take into account the fact that the online content's influence decreases substantially since posting, this consideration is reflected by the exponential rate α. The data fitting results shown in Table 2 confirm that this decay rate is content-specific and varies little from one hop distance to another. This decay rate could be a very important index to measure the value of an online content as a piece of news. The relatively small value of c_0 shows also the insignificant "permanent" influence of the online content. The simulation results shown in Table 2, however, shows some interesting phenomena which deserves further study from networking and social medial point of view: the initial and thus the

Table 2
Parameter values of s1 fitted over distance.

Distance	α	β_0	c_0	σ
1	−46.4156	153.9252	0.7512	0.9346
2	−43.0880	203.0905	1.1492	1.3250
3	−41.2226	270.1044	1.5880	1.7716
4	−39.2731	227.6291	1.6821	1.8679
5	−41.3426	173.5398	1.1590	1.3285

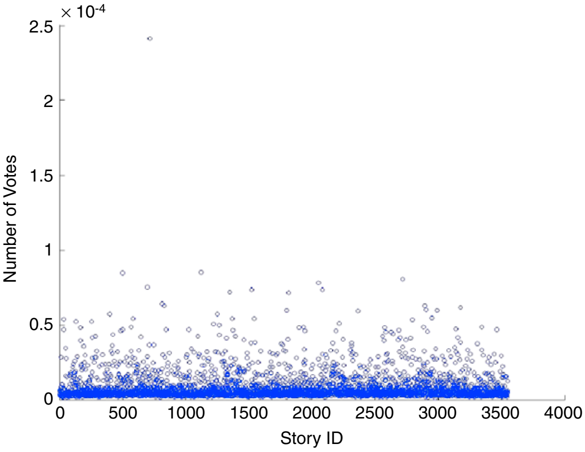


Fig. 3. Number of votes per story, in the dataset including the stories that were promoted to the homepage of Digg.com in June 2009.

maximum influence of the online content increases first and then decreases as a function of the hop distance from the source, so users likely voted for (digg) the content after receiving the content from multiple connected users. A user may not vote for the news the first time she/he receives the news. This observation that multiple exposure to the same content may increase the influence of the content indicates that the information propagation is not completely captured by the diffusion law.

Though s1 provided promising results for the accuracy of our model, the number of votes of s1 in relation to all other voted stories is significantly higher. By examining the scatterplot (Fig. 3), it was clear that s1 can be labeled as an outlier. Results obtained from this story may be indicative of the spread of an extreme piece of online content (or from an epidemiological perspective, a highly infectious virus), and thus extending these results to all other online contents should be examined carefully.

2.3.3. *Small sample results*

The model was fit against the most popular stories, where each had over 3000 cumulative votes by the time of the final vote. Though this sample only accounts for approximately 4% of all stories, the results are significant as they can aid in predicting the dynamics of information spread for forms of online content that are more popular within the DOSN. Fig. 4 shows one particular curve fitting for such a story. Using our model, we achieved an average predictive accuracy of 86.11% for all these stories, thus improving the previous accuracy of Ref. [4] quite significantly. Additionally, we encountered a strong linear pattern between the σ and c_0 parameter values independent of story. This implies that these two parameters change linearly independent of the popular stories chosen.

2.3.4. *Large sample results*

To determine the prediction capability of our model, we fit the model for all stories in the DOSN. On average, the model had an average predictive accuracy of 80.53%. Moreover, in plotting the values of σ and c_0, we obtained a similar strong linear relationship (shown in Fig. 5) of

$$\hat{\sigma} = 1.0089\hat{c_0} + 0.5831$$

thus implying that the difference between the recovery constant σ and the constant rate of infection c_0 is always constant and has no dependence on the stories being analyzed.

3. Conclusion and discussion

By using a variant of the epidemiological SIR model, we obtained an average predictive accuracy of over 98% for the most popular story, approximately 86% for the 134 highest voted news articles and approximately 80% of all stories. These

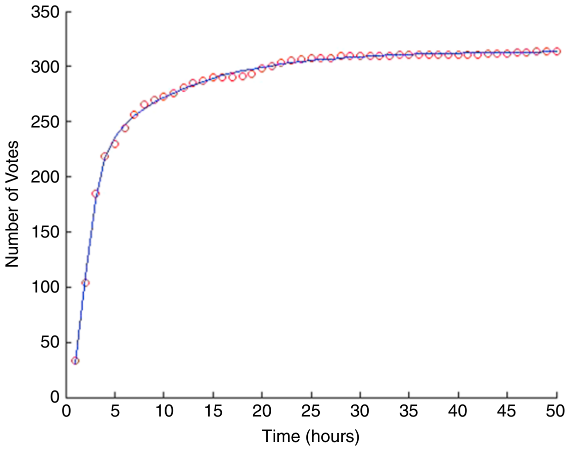


Fig. 4. Curve fit to story 2499.

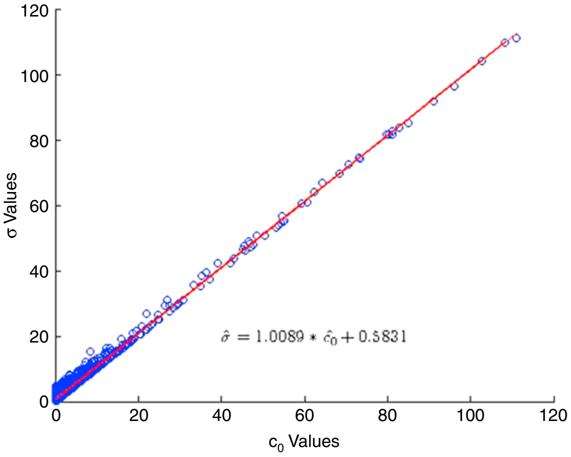


Fig. 5. Linear relationship scatterplot between σ and c_0 for all stories.

accuracies show that the application of a viral information propagation model more accurately predicts the voting trend of Digg network stories than the previous model in Ref. [4] over the first 50 h.

In addition to achieving higher accuracies, we showed that it is possible, given restrictive inequalities, to determine the times at which the number of infected users reaches its maximum, the turning point time in the rate of cumulative number of infected and the beginning and ending times of the viral period for any given story. These properties from this model enable to predict and forecast valuable information for those interested in the spread of media on the DOSN. It will be an interesting study for the future to cluster all stories in terms of some of these critical times to identify story content and network connection features which make these stores share some of these important "epidemiological" moments of viral information outbreaks.

Although we proposed a generic model for the information propagation in online social networks, model parameters are story-specific and hence the model construction is data-driven. As online contents in a network are not really independent from each other, our proposed model should be further expanded to incorporate the "ecological" nature of the competition, cooperation and predation of different online stories in order to examine the online social network ecological system dynamics. This extended model should be in the form of a coupled system with intrinsic dynamics of each story being described by our proposed generic model, much remains to be done to parametrize and analyze such coupled systems.

We have implicitly assumed the homogeneity of users for a news story, and used deterministic version of epidemic models. In the case when the number of infected/influenced users is large, this assumption is likely a good approximation and this also explains why we could achieve nearly perfect fit for the most popular story. When the number of influenced users is small, this assumption of homogeneity is clearly oversimplified and a stochastic version of epidemic models would be most appropriate. This stochastic approach will be feasible if we know the network connection topologies.

We have also used the mass action law (with the simplifying assumption that the susceptible population is sufficiently large while the number of infected users is relatively small) for the *SI*-dynamics, leading to a very simple linear non-autonomous equation for the *I*-equation. We have indeed considered a more realistic equation for the force of infection/influence involving the product of $S(t)$ and $I(t)$. This created some complicated, but manageable, analytic expression for the function of $C(t)$ and $I(t)$ (not reported here). Simulations show that this complication does not lead to any improvement of the data fitting, the parameter identification and key viral indices (turning point, viral periods, peak timing). It would be interesting to see whether the use of standard incidence rather than mass action would lead to any significant improvement. In theory, this would allow us to ignore those contacts between an influenced user and a recovered user. In epidemiology, and specially in dealing with a disease outbreak at a short period of time, the use of standard incidence rather than mass action seems to have led to limited improvement in terms of data fitting. However, the short time scale in the viral spread of an online content in online social networks may require an additional study.

Acknowledgments

This research was supported in part by the Fields Institute for Research in Mathematical Sciences, the Mitacs, the Canada Research Chair program and by Natural Sciences and Engineering Research Council of Canada. The authors would like to thank Professor Feng Wang at Arizona State University for her expertise advice on online social network information dynamics, and Professor Kristina Lerman at the Information Sciences Institute for granting us the access to the datasets of Ref. [1]. We also would like to thank a reviewer for very constructive comments which have led to a much improved manuscript.

References

[1] K. Lerman, Digg 2009 Data Set, http://www.isi.edu/~lerman/downloads/digg2009.html.
[2] F. Wang, H. Wang, K. Xu, Diffusive logistic model towards predicting information diffusion in online social networks, in: Proceedings of IEEE ICDCS Workshop on Peer-to-Peer Computing and Online Social Networking, HOTPOST, 2012.
[3] K. Lerman, R. Ghosh, Information contagion: an empirical study of spread of news on Digg and Twitter social networks, in: Proceedings of 4th International Conference on Weblogs and Social Media, ICWSM, 2010.
[4] F. Wang, H. Wang, K. Xu, J. Wu, X. Jia, Characterizing information diffusion in online social networks with linear diffusive model, in: Proceedings of International Conference on Distributed Computing Systems, ICDCS, 2013.
[5] F. Brauer, P. Driessche, J. Wu, Mathematical Epidemiology, in: Mathematical Biosciences Subseries, 2008.
[6] X. Wang, J. Wu, Y. Yang, Richards model revisited: validation by and application to infection dynamics, J. Theoret. Biol. 313 (2012) 12–19.
[7] S. Tang, N. Blenn, C. Doerr, P. Mieghem, Digging in the Digg social news website, IEEE Trans. Multimedia 13 (5) (2011).