An Epidemiological Approach to Information Propagation in the Digg Online Social Network

Mark Freeman, *Harvard University, USA,* James McVittie, *University of Toronto, Canada,*
Iryna Sivak, *Taras Shevchenko National University, Ukraine,* and Jianhong Wu, *York University, Canada*

***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 show the theoretical properties of this epidemiological model, its applications to social media spread in online social networks and how it more accurately predicts user voting behaviour over a period of 50 hours than the previously established models.**

***Index Terms*—OSN, Digg network**

I. 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, many users will still be influenced by them; however, the causes and dynamics by which information proliferates throughout OSNs are 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 are two forms of connections between users: directed (user A can share content with user B 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.

We used the datasets of [5] which contain 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 24099. 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, [5] 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 [1], 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.

The goal of modelling 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 [2] used a temporal-spatial 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 propogation, they were able to achieve an average model accuracy of 97.41% for the most popular story. Additionally, they obtained an average model accuracy of 97.41% and in analyzing the stories receiving more than 3000 votes (134 stories), approximately 60% of these stories had model accuracies greater than 80%.

Our focus was directed towards an application of an epidemiological model which described the spread of a virus in a population. In using 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 at which the story diffuses quickly through the DOSN and the total time for the story to spread as far as possible in the DOSN. By using this model, we achieved higher model accuracies than [2] in both the most popular story, the most popular 134 stories as well as an average predictive accuracy of approximately 80% for all voted stories.

II. Main

A. The model

We modelled the diffusion of a particular story through the DOSN by using a modified epidemiological SIR model [3]. As in modelling 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. The focus of our study is the cumulative number of people $C(t)$ who have ever voted for a particular story.

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 cumulative rate of spread of the story $\frac{dC}{dt}$ is equal to the product of the number of existing directed network connections from infected to susceptible individuals and the per-connection hourly rate of story spread. Since all the voted Digg stories from [5] eventually became promoted to the Digg homepage but were ultimately not voted for by less than 1.19% on average of the DOSN, then $\frac{dC}{dt}$ 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

$$\frac{dC}{dt} = \beta(t)I(t),$$

where $\beta(t)$ is equal to the per-connection spread rate of the story and $I(t)$ is equal to the average number of outward directed connections per infected individual. We assumed the spread rate exhibited an exponential decay over time t, which yielded

$$\frac{dC}{dt} = (\beta_0 e^{\alpha t} + c_0)I(t),$$

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:

$$\begin{aligned}\frac{dS}{dt} &= -(\beta_0 e^{\alpha t} + c_0)I(t)\\ \frac{dI}{dt} &= (\beta_0 e^{\alpha t} - \delta)I(t)\\ \frac{dR}{dt} &= \sigma I(t),\end{aligned}$$

where $\delta = \sigma - c_0$, σ is a positive constant, and $c_0 < \sigma$. This system has the following set of solutions

$$\begin{aligned}S(t) &= N - C(t)\\ I(t) &= e^{\frac{\beta_0}{\alpha}(e^{\alpha t}-1)-\delta t}\\ R(t) &= \frac{c_0\sigma}{\alpha}e^{-\frac{\beta_0}{\alpha}}(-\frac{\beta_0}{\alpha})^{\frac{\delta}{\alpha}}\Gamma(\frac{-\delta}{\alpha}, -\frac{\beta_0}{\alpha}e^{\alpha t}, -\frac{\beta_0}{\alpha})\\ C(t) &= 1 + \int_0^t (\beta_0 e^{\alpha x} + c_0)e^{\frac{\beta_0}{\alpha}(e^{\alpha x}-1)-\delta x}dx\end{aligned}$$

which equals

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

where N is the number of users in the DOSN and the generalized incomplete gamma function is defined as:

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

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 and $I(t)$ represents the number of "infected" individuals at time t. The relationship between the Richards and the compartmental SIR model from [6] provided the motivation in the application of modelling viral spread to that of the propagation of the story from a particular user (the source) to other users in the DOSN.

B. Properties of the model

1) Long Term Behaviour and Turning Point: In examining the theoretical long term outcome of this model, we denote $C_\infty = \lim_{t\to+\infty} C(t)$. We determined that $C(t)$ asymptotically approaches the following:

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

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 [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 in which 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 $\frac{d^2C}{dt^2} = 0$ and occurs at time:

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

only if

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

2) Virality of a story: We define the "viral period" of a story as the period of time in which the particular story spreads rapidly across the social network. This period begins and ends when $\frac{d^3C}{dt^3} = 0$. The solutions to the cubic equation of $\frac{d^3C}{dt^3} = 0$ have at most two solutions that are real numbers (Proof in appendix). 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 viral period and thus the story does not spread.

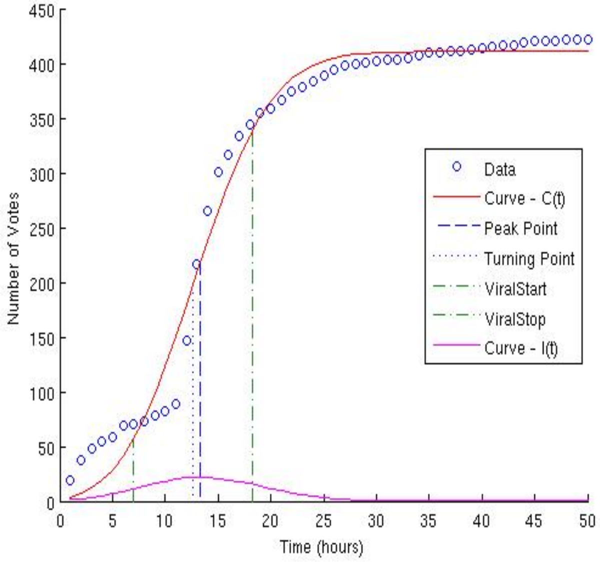

Fig. 1. Visualization of Properties

3) Peak Infection: We define the ”peak infection” time of a story, as the point in time in which $I(t)$ reaches its maximum. This is equivalent to the point in time in which a story has infected its maximum number of users in the DOSN. This value is calculated by solving

$$0 = \frac{dI}{dt},$$

where the peak infection time is denoted by t_{peak} is the following:

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

and occurs only if

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

This restriction ensures that this time is always positive and thus that it exists. If the restriction is not satisfied, then this would imply the situation where a peak infection time would not exist implying that the number of infected users never reaches a maximum thus having $I(t)$ decreasing from $t = 0$.

C. Methods and Results

1) Data Formatting: We analyzed the above model $C(t)$ (integral version) by fitting it to [5]. As in [2], the data was formatted over a period of 50 hours. 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 [1], we used the following measure to determine the model accuracy:

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

where a least squares optimization, with a maximum number of 100000 numerical iterations, was used to determine the four optimal parameter estimates σ, c_0, α and β_0 to minimize $|\text{predicted value} - \text{actual value}|$.

TABLE I
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 %

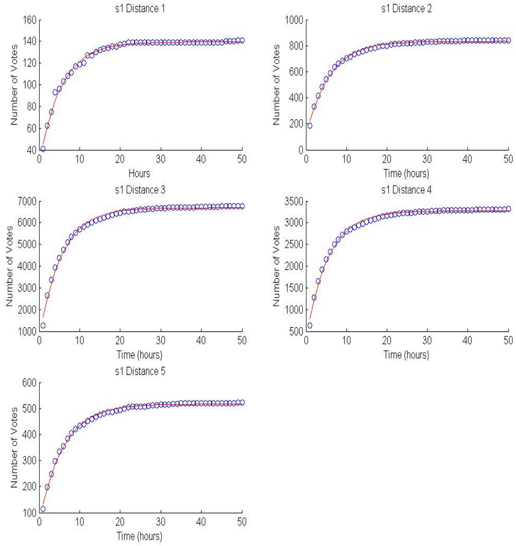

Fig. 2. Fits of Curve to s1 over distances 1-5

2) A case study of the most popular story: We performed a case analysis of the most popular story, 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 [2]. On average, the 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 hours after the initial vote was cast for s1.

Unlike the previous models ([1], [2]) 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 the number of votes for s1 over all 5 distances and plotted this curve along with the points.

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 examinig the scatterplot below (Fig.1), it was clear that s1 can be labelled 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 should be examined carefully.

TABLE II
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

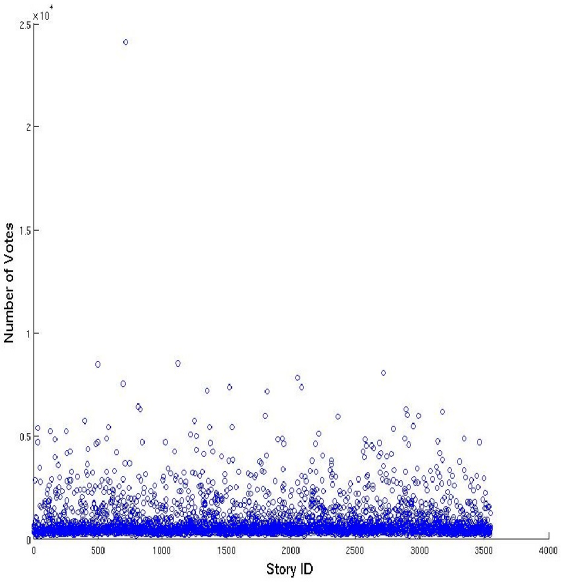

Fig. 3. Number of Votes per Story

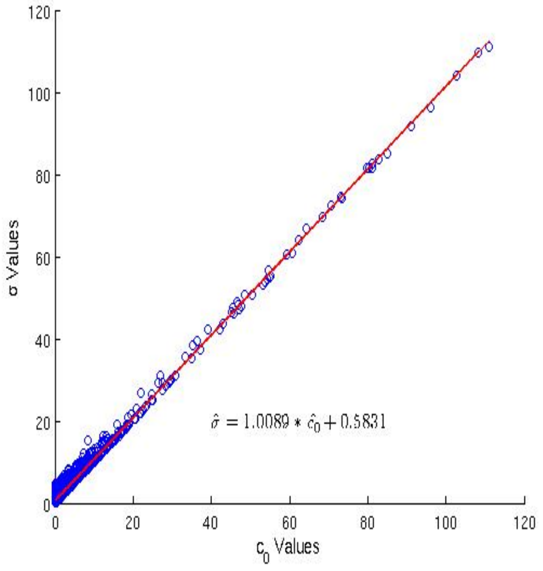

Fig. 4. Linear Relationship Scatterplot between σ and c_0 for all stories

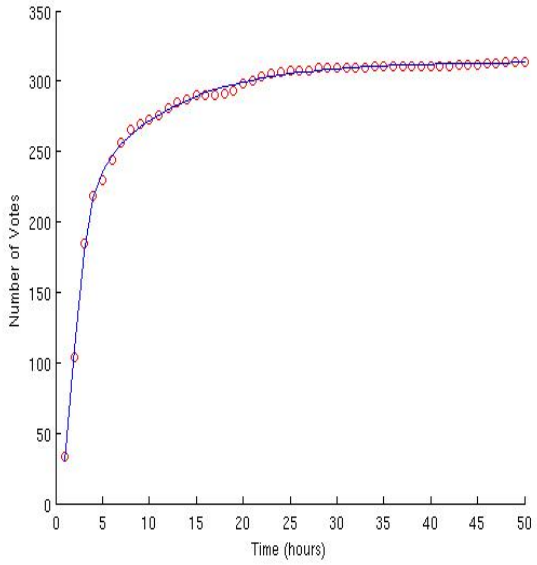

Fig. 5. Curve Fit to Story 2499

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. Using our model, we achieved an average predictive accuracy of 86.11% for all these stories, thus improving the previous accuracy of [2]. 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.

4) Large Sample Results: To determine the prediction capability of our model, the model was fit 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 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.

III. 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 stories and approximately 80% of all stories. These accuracies show that the application of an epidemiological model more accurately predicts the voting trend of Digg network stories than the previous model in [2] over the first 50 hours. 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. By using these properties from this model, predictions and forecasts can be made which will predict valuable information for those interested in the spread of media on the DOSN.

Appendix A
Proof of $t_{peak} > t_{turn}$

We will show that the time at which there is the maximum number of infected t_{peak} is always greater than the turning point of $C(t)$ denoted by t_{turn}.

$$\frac{dC}{dt} = \beta(t)I(t)$$

then

$$\frac{d^2C}{dt^2}(t_{peak}) = \beta'(t_{peak})I(t_{peak}).$$

$$\frac{d^2C}{dt^2}(t_{peak}) = \alpha(\beta_0 e^{\alpha t_{peak}} I(t_{peak}))$$
$$= -\alpha(c_0 - \sigma)I(t_{peak}) < 0$$

Because after the turning point, $\frac{d^2C}{dt^2}(t_{peak}) < 0$ which implies that the peak point is always greater than the turning point. QED

Appendix B
Proof of two positive, real solutions of $\frac{d^3C}{dt^3}$

We will show that there are most two positive real solutions of $\frac{d^3C}{dt^3}$. By taking the third derivative, a cubic equation was obtained.

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

where $B = \beta_0 e^{\alpha t}$. Notice that the leading coefficient and constant terms are both positive. Then

$$\lim_{B\to-\infty} \phi(B) = -\infty$$

and

$$\phi(0) > 0.$$

Then by the intermediate value theorem, since ϕ is a continuous function of B then it must have at least one negative zero. This implies the existence of at least one complex time solution which implies the existence of at most two real time solutions. QED

Acknowledgments

We would like to thank the Fields Institute and MITACS for all of its support in funding our research.

References

[1] F. Wang, H. Wang, and K. Xu, *Diffusive logistic model towards predicting information diffusion in online social networks*, Proceedings of IEEE ICDCS Workshop on Peer-to-Peer Computing and Online Social Networking (HOTPOST), 2012.

[2] F. Wang, H. Wang, K. Xu, J. Wu, and X. Jia, *Characterizing Information Diffusion in Online Social Networks with Linear Diffusive Model*, Proceedings of International Conference on Distributed Computing Systems (ICDCS), 2013.

[3] F. Brauer, P. Driessche, and J. Wu,*Mathematical Epidemiology*, Mathematical Biosciences Subseries, 2008.

[4] S. Tang, N. Blenn, C. Doerr, and P. Mieghem,*Digging in the Digg Social News Website*, IEEE Transactions on multimedia, Vol.13, No.5, 2011.

[5] K. Lerman, *Digg 2009 data set*, http://www.isi.edu/ lerman/downloads/digg2009.html.

[6] X. Wang, J. Wu, and Y. Yang, *Richards Model Revisited: Validation by and application to infection dynamics*, Journal of Theoretical Biology, 313, 12-19, 2012.