How Does News Diffuse Through Twitter?

Predicting spread of information through social media using a Mathematical Model

Caleb Davis, Sandra Ramirez, Daniel Whitmore

Why?

Knowing how quickly information spreads across the Internet is important to many fields. Advertisement could be optimized by choosing the best possible 'source'. The speed in which people communicate has a profound social impact. Using these models, the speed, and thus impact, of a particular message being spread from person to person can be predicted.

Barack Obama tweeted this photo. The rate at which this tweet spread will be used as an example data set.

Twitter

This mathematical model is analyzing information flow within a social networking website. Twitter provides vast databases of actual rates of information flow over time and how many people are participating. Within the twitter social network, users “follow” other people with twitter accounts. These users can follow their friends, celebrities, or even famous politicians. By being a “follower”, one can view the tweets, and also, “retweet” a person’s message. When a person “retweets” a status or a picture, he or she is reposting the tweet so that his or her followers can now view the tweet. By retweeting, followers are practicing information diffusion through online social networking.

In this experiment, we seek to explore how a Partial Differential Equation (PDE) can model how information is diffused throughout the world. By doing so, we look at the density of the population that is retweeting and the distance between the followers that are retweeting certain messages. The information acquired from this data can be very useful in determining the information diffusion process in both temporal and spatial dimensions in online social networks.

Mathematical Model

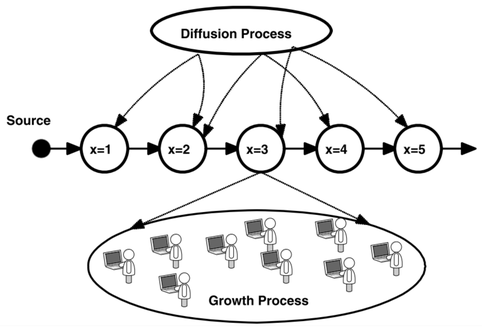
The Diffusive Logistic (DL) equation including boundary and initial conditions for modeling Information Diffusion:

$$\frac{\partial I}{\partial t} = d\frac{\partial^2 I}{\partial x^2} + rI\left(h(x) - \frac{I}{K}\right)$$

$$I(x,I) = \varphi(x),\ I<x<L$$

$$\frac{\partial I}{\partial x}(I,t) = \frac{\partial I}{\partial x}(L,t) = 0, t>1$$

- I(x,t)- dependent variable that represents the density of influenced users with distance x at time t (x and t are independent variables);
- d is a constant representing the social capability (a measurement of how fast information travels across distances in social networks);
- r(t) is a function that represents the intrinsic growth rate of influenced users at the same distance (measures how fast information diffuses amongst users at the same distance);
- h(x) is a function that adjusts the density of influenced users within a certain group allowing each group's density at distance x to be adjusted independently of other groups' densities at other distances;
- K represents the carrying capacity meaning the maximum possible density of influenced users at a given distance;
- *l* and L represent the lower and upper bounds of the distance from the source and other twitter users. (x = 1 is the source's friends; x = 2 is friends of x=1, but not the source; x=3 is friends of x=2, but not x=1 or the source)



This model is believed to be the first of its kind to model the temporal and spatial characteristics of an information diffusion process in online social networks by way of a nonlinear PDE, the Diffusive Logistic equation.

It is important to understand that this is a discrete process (each group has a discrete distance from the source) being modeled with a continuous curve model (using cubic splines) and that after modeling thousands of data sets, Wang and associated researchers have achieved higher than an 80% average accuracy rate via the DL equation.

The Real Data

Using MATLAB we were able to run tests on the twitter data provided in an attempt examine how online social networking can be modeled by a diffusion equation. By adjusting factors in the growth function, the t_{max} values, and the h(x) function, we obtained several plots of the model. With these plots, we were able to observe how a variable h(x) function resulted in high accuracy. We chose to display our most accurate results, which can be viewed in the following figures.

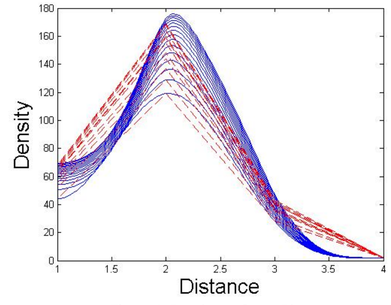


Figure 1: Density vs. Group Distance
T_{max} = 15 hours, accuracy=.9764

Results:
The mathematical model represented the diffusion of President Obama's tweet to an accuracy of 97.64%, shown in the figure above The red lines represent the actual number of people participating in the news at various time increments. The blue curves represent the model used to predict the information diffusion. These results were obtained with a variable h(x) function and a t_{max} value of 15 hours. The growth rate function in this particular case was: $r(t) = 0.3 + e^{(-2t)}$. Figure 1 illustrates a peak at x=2, which can represent how the popularity of President Obama’s tweeted picture decreased at distance two.

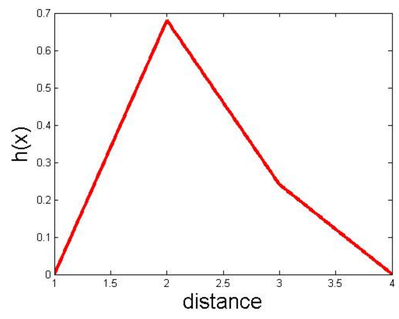


Figure 2: Social distance function h(x) vs. distance from source

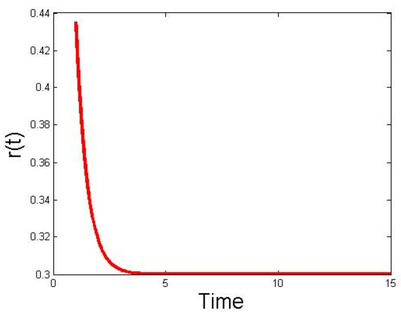


Figure 3: Growth function r(t) for tmax=15

Conclusion:
Using this mathematical model, the rate at which news spreads from person to person can be predicted. This same model can be used to find which source is the most effective for spreading information and can be used to predict the rate at which current news is spreading.

Acknowledgments: Feng Wang, Haiyan Wang, and Kuai Xu initiated the development of the PDE models. The project is supported by a grant from NSF.

References: Diffusive Logistic Model Towards Predicting Information Diffusion in Online Social Networks (F. Wang, H. Wang and K. Xu), 2012 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW), 2012, pp.133-139