Geo-Based Twitter Flu Data Analysis and Modeling

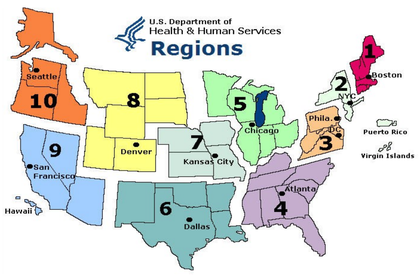
Utilizing Logistic Partial Differential Equations to Model Health Related Information spread in Twitter

Kyle Brown, Andrew Cannella, Jaime Chon, Dustin Leffew, Ross Raymond

The Purpose

The influence of social media is continuing to permeate our culture at an ever increasing rate. Using a mathematical model, we seek to more efficiently track the spread of flu-related media. The model we have created to effectively predict how the information will spread through social media, particularly through Twitter, uses a logistic partial differential equation. Last year's presentation of this project detailed the growth of research to modeling information diffusion for the spread of flu. This time, we compare our data collected to actual CDC records to predict correlations between Tweets and real occurrences of the flu.

Tweet on the flu gathered from CNN's Twitter account. Examples of keywords are highlighted.



CDC breakdown of United States into Regions for ILI (influenza-like illness)

Data Gathering

The Twitter data was gathered from Twitter from January 3rd to March 26th of 2014, using a combination of Twitter API calls and python coding. We query Twitter to grab tweets containing particular key words, which we then iterate over to gather unique user IDs, locations, and source tweets. Replies and retweets are then recursively mapped onto their source tweets, allowing us to compare the IDs of each reply and retweet to the ID of the source user. Previously, we used this mapping to add users to distinct friendship distances to measure the density at each level. For this project, we group users based on their location using the 10 regions enumerated by the CDC's "Overview of Influenza Surveillance in the United States."

Finally, each region is given a count of unique users and placed into a matrix of Region vs. Time. Time intervals chosen extend over a 1-week period and are cumulative. The matrix is printed to file and used within our MATLAB simulations. At the same time, we collect data from the CDC's record of flu reports and run simulations on the sets concurrently.

Architecture and Mathematical Model

In this project, we utilize a five-layer architecture for modeling and analyzing information diffusion in online social networks that build on each other: 1) online data collection; 2) cyber-distance definition in online social networks; 3) Partial Differential Equations (PDE) model development; 4) PDE analysis; and 5) applications. We continued the research of our previous projects and focus on clustering, PDE analysis and application.

5. Application
4. Model Analysis
3. PDE Model Development
2. Define cyber-distance
1. Data Collection

The main difficulty in creating a model to predict information diffusion within social media stems from the complexity of how that information actually spreads. Typically, information follows either a content-based or a structure-based diffusion process. A content-based process depends upon the immediacy and presentation of the information, taking on a more isolated occurrence. A structure-based process is similar to when a topic is posted by a source, accessed by the source's followers, and retweeted to the followers of the followers. It describes how information spreads through each intermediate point.

The following Diffusive Logic (DL) equation was derived to model these different processes:

$$\frac{\partial I}{\partial t} = \underbrace{d\frac{\partial^2 I}{\partial x^2}}_{(1)} + \underbrace{rI\left(h(x)-\frac{I}{K}\right)}_{(2)}$$

(1) Content-Based Diffusion
- ***d*** – constant that measures the speed at which information spreads across distances

(2) Structure-Based Diffusion
- ***r(t)*** – function that measures diffusion speed in the same distance
- ***I(x,t)*** – dependent variable representing density at distance ***x*** and time ***t***
- ***h(x)*** – function to adjust density for each distance independently
- ***K*** – carrying capacity of the density at each distance

The authors extend their gratitude to the NCUIRE program.

New College Undergraduate Inquiry and Research Experiences NCUIRE

Data and Simulation

We used MATLAB code to simulate the data via our mathematical model to the left. We manually adjusted the t_{max} values and the h(x) function to increase accuracy. MATLAB then printed out the accuracy at each distance, the overall accuracy, and several plots, the primary of which is shown below for both the Twitter data and the CDC data.

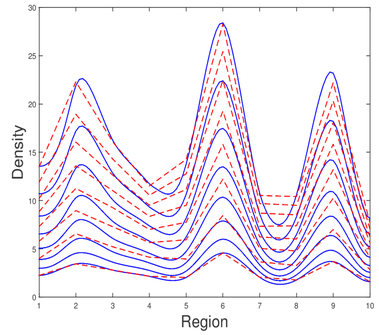


Figure 1: CDC Flu, Region vs. Density
T_{max} = 8 weeks, Accuracy = .767

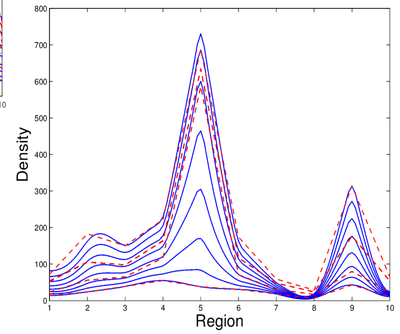


Figure 2: Twitter Flu, Region vs. Density
T_{max} = 8 weeks, Accuracy = .721

Results:

Figure 1 is a plot of Region vs. Density for the CDC data; Figure 2 for our gathered Twitter data. The red lines correspond to the actual count of reported cases or users at each Region connected linearly, while the blue curves correspond to the simulation based on our formula. As we can see, there are peaks at x = 2 and x = 6 in both figures. However, around Regions 5 and 6 there is a discrepancy in the peaks. This could correspond to a difference in regional densities, suggesting that while Region 6 experienced more reported cases of influenza, Region 5 had a higher rate of flu-related tweets.

Conclusion:

A diffusive logic partial differential equation can accurately predict the spread of flu-related information through Twitter. Comparison between mediums for further prediction can be difficult given the variant nature of their data, but does suggest this equation can be used to appropriately model data from other sources.

Acknowledgements:
This project is directed by Dr. Haiyan Wang and Dr. Feng Wang, and is supported by Kuai Xu and a grant from the NSF.

Reference:

Modeling Information Diffusion in Online Social Networks with Partial Differential Equations (F. Wang, H. Wang and K. Xu).
Overview of Influenza Surveillance in the United States (CDC).