Proposal

Quentin Vilchez, Fin Vermehr and Binderiya Adishaa

March 27, 2018

1 Goal:

- Our goal is to better understand how news spreads on a social network, in our case we wish to focus on Twitter.
  We would like to apply a compartment model similar to the SIR model.

- Our question would be:
  How does the existence of a susceptible population, with a history of infection by similar topics, modify the outbreak of a story on a social media?

2 Our Approach:

- There has been some work done on this topic, specifically by Bettencourt et al. [1], who developped the SEIZ (Susceptibles Exposed Infectives Skeptics) model to describe the adoption of the Feynman diagrams by the scientific community around the world. This model introduced an exposed state and skeptic state, individuals in the exposed state take some time before they begin to believe/react to a story, individuals in the skeptics state do not believe the information and will not react to it. Furthermore, Fang Jin et al. [2] used this model to describe how information propagates on Twitter and demonstrated how this model is able to accurately capture the patterns in the spread of new/rumors on Twitter.

- We would like to base ourselves on the work of these individuals, and in a similar manner to Fang Jin et al. [2] describe how information spreads on Twitter. However, we think that it could be interesting to make a change to the model, by introducing a second susceptible population, which would account for people who have a history of tweeting about topics similar to the one in question. For example, if we study how the #parkdaleshooting propagates on Twitter, similar topics would be #guncontrol, #NRA, etc. So people in the second susceptible population would be people who have tweeted about these topics in the past.
  The reason for introducing this second population is that we feel that this second population has a much more significant influence on the virality of certain content. Meaning that a story might become viral in a shorter span of time if there is a large second population and also they might cause other surrounding topics to go viral again.
  That is why we would like to focus on two different kinds of stories, one with a large population that is more susceptible than usual and one with a small population that is more susceptible than usual. We believe, that the outbreaks will be different and that this can be described by

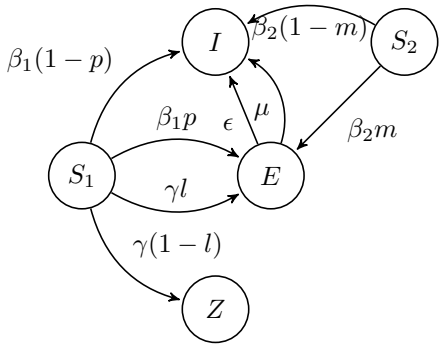
our model. We will also compare our model to the simple SEIZ, to determine whether ours can fit the data better.

- These are the equations that will describe our system:

$$\begin{cases} \dot{S_1} &= -\beta_1 S_1 \frac{I}{N} - \gamma S_1 \frac{Z}{N} \\ \dot{S_2} &= -\beta_2 S_2 \frac{I}{N} \\ \dot{I} &= \beta_1(1-p) S_1 \frac{I}{N} + \beta_2(1-m) S_2 \frac{I}{N} + \epsilon E + \mu E \frac{I}{N} \\ \dot{E} &= \beta_1 p S_1 \frac{I}{N} + \beta_2 m S_2 \frac{I}{N} + \gamma l S_1 \frac{Z}{N} - \mu E \frac{I}{N} - \epsilon E \\ \dot{Z} &= \gamma(1-l) S_1 \frac{Z}{N} \end{cases}$$

with,

$$N = S_1 + S_2 + E + Z + I \qquad \beta_1 < \beta_2 \qquad 0 \leq m < p \leq 1$$
$$0 \leq l \leq 1 \qquad \dot{N} = 0$$



3 Analysis and Data Fitting:

- We will first do an analysis of the system of equation by looking at the equilibrium and their stabilities.
- We have been gathering large amounts of data from Twitter. We will use this data in order to:
  1. Create hashtag clusterings. These clusterings will allow us to see and understand what topics are close in nature, to the topic that we are trying to describe. These clusterings will also enable us to determine an approximate size for the second susceptible population at $t = 0$ and hence the size of the first susceptible population at $t = 0$ as well.
     To cluster hashtags we have developped two algorithms based on the work of Ali Javed [3] on semantic hashtag clustering in social media.

2. We will estimate unknown parameters by fitting the equations to the data. For that we will follow what Fang Jin et al. [2] did except that we are going to try to estimate initial susceptible population sizes. The data that will be used for fitting is the cumulative number of tweets with respect to time.

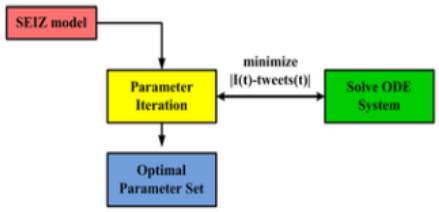


Figure 1: Numerical implementation work-flow.

4 Expected Results

We expect to see outbreaks that happen earlier, are more intense and last longer when the second population is large enough. Given that the second population will have a positive feedback on the first one, meaning that the second population is more likely to get infected, the number of infectives will climb rapidly and therefore more and more individuals from the first population will be infected.
We believe that the introduction of a second, more susceptible, population will enable us to accurately predict for a given story on twitter its steady state infected population size and the speed at which it reaches its equilibrium state.

5 References

[1] L. Bettencourt, A. Cintron-Arias, D. I. Kaiser, and C. Castillo-Chavez. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. PHYSICA A, 364:513–536, 2006.

[2] F. Jin, E. Dougherty, P. Saraf, Y. Cao, N. Ramakrishnan. Epidemiological Modeling of News and Rumors on Twitter, SNAKDD '13 Proceedings of the 7th Workshop on Social Network Mining and Analysis Article No. 8.

[3] A. Javed. A Hybrid Approach to Semantic Hashtag Clustering in Social Media, Graduate College Dissertations and theses, 2016.

[4] J.M Heffernan, R.J Smith, L.M Wahl. Perspectives on the basic reproductive ratio, Published 22 September 2005.DOI: 10.1098/rsif.2005.0042.