About the Project

D. Brown, B. Fullwood, N. Konz, T. Richards, E. Yelton

October 7, 2018

1 Main Idea

The questions that sparked this project were 'How are ideas spread?' and 'Can we model this behavior?'. With internet access, "ideas" are spread more rapidly than ever. Under the current polarized political climate as well as controversies of foul play by social media platforms there seems to be a heightened interest in such behavior. Our hypothesis is that **the spread of ideas can be modeled by the behavior of infection.**

2 Method

To probe this hypothesis we used data retrieved from Reddit, using the python Reddit API wrapper PRAW. We used this data to get a measure of the 'connectedness' L between various subreddits; This term is a metric of how many users on one subreddit are also active on some other chosen subreddit. For example, if we want to determined the connectedness L of two subreddits indexed with j and k, we look at the top $M = 100$ posts in the smaller (fewer subscriber) subreddit, and see how many of the authors from that list have posted in the other subreddit. So, we define L as

$$L_{jk} \equiv \frac{N_{\mathrm{j,k}}}{M} \equiv \frac{\text{No. of users that post in both}}{\text{Total no. of users sampled}}. \tag{1}$$

These L values are part of the realistic basis for our simulation.

The mathematical model we used to simulate the spread of the ideas within a single subreddit is derivde from the SIR model of the spread of disease [1], which gives a set a 3 coupled ODEs described by D. Smith and L. Moore,

$$\begin{aligned} \frac{ds(t)}{dt} &= -bs(t)i(t) \\ \frac{di(t)}{dt} &= -bs(t)i(t) - ki(t) \\ \frac{dr(t)}{dt} &= ki(t) \\ \frac{di(t)}{dt} + \frac{ds(t)}{dt} + \frac{dr(t)}{dt} &= 0 \end{aligned} \tag{2}$$

where the s, i, r are just the percentages of the population (of a given community) susceptible, infected, and recovering, respectively (mutually exclusive of course). In our analysis we used the analogy that s is the percentage of users of the given subreddit who have not seen the 'idea', i is the percentage of users actively discussing and/or posting about the idea, and r corresponds to the percentage of users who have previously been 'infected' and are no longer posting on such an idea. Note that the above equation have constants b and k that correspond to the contagion constant (the number of 'contacts' needed to spread the idea) and the recovery constant (the fraction of people who will forget about certain idea over a certain time interval), respectively. The last equation indicates that our model assumes that the total population of Reddit users for a given subreddit does not change over time. Note that we also assume the constants b and k to be different for each subreddit or community. This means that an idea's propagation will depend on the subreddit community that it's receiving it. For example, subreddits that generally facilate the fast spread of new ideas will have a higher b, while subreddits that usually forget about popular trends quickly will have a higher r. For our simulations, we took educated guesses about the best b and r values for the subreddits.

In order to actually solve the ODE's for each community that we're considering, we used a Runge-Kutta 4th order integration method (ODE solver) to solve the system of equations, which we run for each time step, for each system of each community.

However, these ODE's only model the spread of ideas **within** subreddits. We also model the spread of ideas **between** subreddits, using a form of conditional probabilistic Markov chain derived from the connectedness values. Say that we have N subreddits in total, each with different L connectedness values between every pair of them. For every noninfected subreddit, at each time step we estimate the probability of the subreddit becoming infected from any one of the infected subreddits, using the following (normalized) estimate (which sums over the other $N-1$ subreddits, which are indexed by k:

$$\text{Probability for infection of } j^{th} \text{ subreddit} \equiv P(t)_j = \frac{1}{N-1} \sum_{k \neq j}^{N-1} L_{jk} i_k(t) \quad (3)$$

Notice that the $N \times N$ matrix of L values is essentially used as a transition matrix from Markov chain theory. The higher this probability, the higher the chance that at the time step, the jth subreddit will become infected. This calculation is done for each of the uninfected subreddits.

Given more time than the 24-hour limit, we would've liked to slightly correct the infection probability computation by accounting for the generalized inclusion-exclusion principle of probabilities (to avoid over-counting). We also wanted to compare our results to real reddit data over time, but it was difficult to get this type of specific data for the precise range of times that we needed. We visualized this entire process on the main page.

References

[1] Smith D., Moore L. (2004 December) "The SIR Model for Spread of Disease - The Differential Equation Model" *Convergence*
Retrieved October 6, 2018 https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model