Michael Maurer and Lauren Winston
Project 3 Proposal

Literature Review

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC128598/

In the article, the author addresses pros and cons of an agent based model in relation to simulating human systems. One of the biggest take aways that influenced our future research was his statement that: “a set of differential equations, each describing the dynamics of one of the system's constituent units, is an agent-based model.” At the time of reading this we were considering the different modeling techniques that we could use in our goal of simulating a social network. His argument for agent-based modeling was convincing as he described them in terms of many decision making entities. This almost perfectly fits our definition of a social network. His cons of agent-based modeling also influenced us. He states: “Another issue has to do with the very nature of the systems one is modeling with ABM in the social sciences: they most often involve human agents, with potentially irrational behavior, subjective choices, and complex psychology—in other words, soft factors, difficult to quantify, calibrate, and sometimes justify.” This led us to other areas of research over different agents in a social network. As seen below in other papers, although each individual is highly unpredictable, statisticians have found that general network behavior can be predicted.

http://www.nature.com/srep/2012/120329/srep00335/full/srep00335.html

In this paper a study of an agent-based model for social network simulation is conducted and described. The analysis of this paper was conducted primarily in relation to Twitter, but the models used can be extrapolated to other social networks, such as Facebook. There are inherent limitations and nuances in which social network is analyzed. Facebook is more limited than Twitter because there is a more limited set of publicly accessible data. However, Facebook’s social network is more closed than Twitter’s so ideas and content spread much differently than through Twitter.

Similarly, the purpose of the paper was primarily to study how meme lifetime and user interests affect competition among memes, but the described model and concepts for social networks can be applied to various other types of simulations.

The model used by this study included three parameters: how focused or diverse user interests were across the network, how often a post is re-shared, and how much new information or ideas users come across outside of the social network. Each user on a social network has a screen by which new ideas can be found and memory by which ideas can be stored. Non-actions and actions that a user can take when interacting with social media form sequences with specifically associate probabilities.

http://crpit.com/confpapers/CRPITV61Ahmad.pdf

This conference paper research meme spread over a social network of IM users. It mathematically described and analyzed. They described the simulation almost as an infection scenario. The meme started with one person and spread its way through the network. The assigned the probability of transmission as a conditional probability depending upon the set of paths between vi and vj. They also built in the case where a user was offline by temporarily removing it from the network. The most interesting part of this paper was the conclusion. It states: “It was discovered that not only the connectivity of some of the nodes (hubs) determine how fast the meme is proliferated but also the time-span in which the corresponding person is online or offline.” This led us to strongly take into consideration how probability of a user seeing a meme is determined. It also made us think about how the most popular people (hubs) affect the network and what it means to be popular.

Conceptual Model

1) Problem that we’re modeling:
    a) The problem that we’re trying to solve is showing people how ideas spread on social media. Since the main intent is education, we plan on presenting a lot of unique informative information to the user.
    b) Since the intent is education, as we will let users control the parameters to see how changes will impact the spread of ideas, however, we will also try to use past research to determine good values for transmission of information and other probabilities such as likelihood of observing a meme based on the total amount of shares from connected nodes.
    c) In addition to information spread, we want to provide users with different terms in relation to social networks and how their own network compares. For instance, we found an interesting term called “broker”. A broker in general is someone who possesses strong ability to transmit ideas in terms of a network. One measure of a broker is to take the shortest path between all points. The number of occurrences for each node is a measure of brokerage.
    d) It might also be interesting to see how user social networks compare to other data occurring in nature. For this, we can fit network data to the power law and explain the concept of the power law to our users. For

example, the degree of nodes in a graph is said to follow the power law often in real social networks..

2) Who the customer of the model / simulation is
    a) People who wish to learn about the spread of ideas and content on social media. The most exciting part of our simulation is that it can theoretically be used by anyone with their own social network (at least in the case of Facebook). Seeing who the biggest hubs of information are and who the information brokers are is a fascinating exercise for a lot of people. For others, we suspect that seeing how their network may closely follow natural laws (power law) would also be very interesting.
3) How this relates to prior work (from the literature review)
    a) Our simulations will take and build from a lot of the concepts and models of earlier related works. However, these simulations are not widely available online in a way that can be manipulated by users. Our simulation will live as a web app that will let any visitor immediately test the simulation and manipulate parameters. However, we also seek to offer comparisons of how ideas spread on different social networks.
        i) Twitter - Retweets of tweets
        ii) Facebook - Shares of posts
        iii) Reddit - “Upvote-ability” of a post - How the biggest sharers influence the most popularity of posts
        iv) Pinterest - Repins
    b) Our model will be based specifically on the model described in Weng’s study involving the different actions of a user when using social media, based on a sequence of probabilities.
        i) There are two key actions that can occur when a user is browsing a social network site: new content is shared with the users followers or previously shared content is re-shared by the user.
        ii) There is a certain probability, p, of someone posting something new on social media and a corresponding 1 - p probability of scanning the screen.
        iii) Each post on the screen has a probability of catching the user’s attention, p_r. Of the posts that catch the user’s attention, there is another probability p_m that the post triggers the memory of another post and the corresponding 1 - p_m probability of the user re-posting the content that caught his attention.
        iv) Each of these probabilities will be in some form available as manipulatable parameters in our simulation.

Implementation Plan

Languages:

We will mainly use JavaScript to build the web app. We will use PHP for our server needs. We will use a JavaScript data visualization package to graphically represent our data.

Sources of Data:

We will use publicly available data in addition to more private data from our own social networks through public APIs for the following social networks: Facebook, Twitter, Reddit, and Pinterest.

Our intent is to directly compare at least two of the four social networks listed. After initial attempts to access the data, we will have a better idea of the feasibility of getting data from each social network. We anticipate issues with accessing APIs and with the usefulness of the information we are able to get from the APIs. Analyzing four separate social networks is out of scope for this project given our time restraints, so we will select two or three social networks to proceed with for the rest of the simulation based on the quantity and quality of data we are able to access.

Describe Simulation Experiments:

Using the APIs we will get real data from the different social networks related to shared content. We will use this data to re-construct models for each social network. This data will be used to form probabilities of the described in the conceptual model.

In the actual network simulation, we can start with an initial state and work for some amount of time. We want are simulation to be in semi-real time so we can visualize it. The problem that we may run into is that the size of the network may prevent us from calculating and rendering in real time. In order to combat this, we have two options: create a real-time timestep representing some amount of time in real life. Between time steps, we can calculate the next state and update it at the end of the time step. Another option would be to run the simulation before visualization and display the results closer to real time after all the computing is done. Javascript may be limiting in this regard, so limiting or cutting down the size of the network may be necessary. Over time, we can compute statistics to show the user who was the most responsible for information spread in their network under the current parameters. Other statistics include the total amount of network saturation and the effects of removing nodes from the network.

Simulations Input and Output

- Input

    - Which social network
    - Double value representing diversity among interests in network groups
    - Double value representing probability that a user will post new, unique content to a social network
    - Double value representing probability that a user will re-share content that interests him or her
    - Network size
- Output
    - Visual representation of the spread of a new meme generated at the source. This is essentially a directed graph with different nodes representing users. Different colors will be used to represent the source of the post, nodes that see that post at some point throughout its lifetime, and nodes that share the post at some point throughout its lifetime.
    - A numerical summary corresponding to the visual representation, including the number of nodes the post passes through and how many levels of the network the post passes through.

Goals for Checkpoint C:

- For this checkpoint we are going to focus on having access to the APIs. Our goal would be to have some form of basic data, with the ability of accessing more, for at least two out of the four social networks described.

Goals for Checkpoint D:

- Our goal for the end of this checkpoint is to have the framework for the visualization of the web app implemented.
- Analyze the data from the social networks and work on incorporating this data into our model.
- Numerical (non-visual) output should be produced.

Goals for Checkpoint E:

- Functional simulation on web app for one of the social networks.
- Visual output (network of nodes and edges) should be produced with each simulation trial.