Analyzing Information Flow Through Stochastic Processes

Team 48736

Contents

1 Introduction

Throughout history, humans have always had methods of sharing new and interesting information with each other. While people thousands of years ago found themselves limited to stories relevant to their local communities with details of the world coming only from passing travelers, today we can use our phones to check current global events in a matter of seconds.

Throughout United States history specifically, there have been five time periods where people used different mediums as their main sources of information. The five mediums and periods are as follows:

- **Newspaper and telegraphy in the late 1800s** forced local news to stay local, with only the biggest stories reaching a national scale through telegraphy.
- **Radio in the 1920s and 1930s** allowed information to reach people faster because news no longer had to be published; broadcasters could just announce events over the air as they happened. National stations also unified the country by broadcasting the same stories everywhere.
- **Television in the 1960s-1980s** easily took over radio because people felt more engaged seeing and hearing news, as opposed to just hearing it.
- **Early internet in the 1990s** gave people the choice to check news at their own leisure, as opposed to waiting for a TV broadcast.
- **Smartphones in the 2010s** now provide people the ability to access any information they want wherever they go.

As part of an effort to analyze the evolution of society's information networks, these time periods serve as clear dividing lines between different ways Americans have communicated information with each other across history.

1.1 Approach to the Problem

In this paper we present a computational model to analyze the rate at which information of different inherent values from different mediums spreads across networks of people over time and to determine what information qualifies as "news". We do this by focusing specifically on the five periods in United States history mentioned above.

We validate the results of our model by comparing them to known information trends in recent history from Google Trends. We also use our model to predict how information will spread across networks of people in the year 2050. Additionally, we explore how people's initial opinion or bias on a topic, the medium of information, and the inherent value of information can be used to influence public opinion.

1.2 Assumptions

Due to the difficulty of quantifying the inherent value of some piece of information, the number of factors that actually affect how information spreads in real life, and the lack of concrete historical data, we make a few assumptions in order to simplify our model and maximize its potential accuracy:

- **Information is equally compelling no matter where it is.** That is, an interesting story will be interesting regardless of its origin.
- **Groups of people receive information.** Rather than treating information as something spread between individuals, we treat it as something that is spread between groups of people.

- **Groups of people are connected.** Any individual in a group of people must know at least one other person in another group of people; therefore, all groups of people in our model have connections to at least one other group.
- **Groups of people share information with each other.** If a group of people hears or sees new and interesting information, they will share that information with another group they are connected to. This continues ad infinitum until the information is no longer interesting.
- **The major population centers of the United States have stayed the same over time.** Although the overall United States population has increased since 1870, the majority of the population has stayed concentrated in the south and northeast.
- **People who have access to a news medium will see stories that come to that medium.** While someone owning a TV in their home does not necessarily guarantee they will use that TV to watch news, we assume simply having access to an information medium means getting information from that medium.
- **Today information spreads similarly despite geographical differences.** Numerous socioeconomic and political factors have shaped people's access to information in different countries throughout time. Because of this, we specifically focused our modeling of historical time periods on the United States to avoid inaccurate generalities. Due to the current global nature of the internet, however, we think it is safe to assume interesting information spreads across most countries the same way it spreads across the United States.
- **As the number of available informational mediums increases, people use all available mediums.** That is, we assume people in the 2010s get their information from newspapers, radios, TV, internet, and smartphones combined.
- **People trust information they have heard before more than completely new information.** If a group of people hears a piece of new information from another group of people, they may not immediately consider that new information true; depending on the source, it might just be a rumor. In our model, therefore, the probability of a group believing new information increases the more times that group has had the information shared with them.

2 Existing Models

Several models currently exist which can be used to model either information networks or, more generally, stochastic processes. Our model uses elements from two in particular: the independent cascade model and Poisson processes. We also use Monte-Carlo simulations to validate that the performance of our randomized model is not accidental.

2.1 Independent Cascade Model

The independent cascade model for information flow models a social network as a graph. J.R. Lee and C.W. Chung introduced the approach in 2012. [**?**] In this model, when a vertex in a graph learns new information at any given time t, it has a chance to spread that information to its neighbors with some probability p, but it only has one chance to do so. After this, a time step is taken and the process continues until all the vertices in the graph have either learnt and passed on new information, learnt and not passed on new information, or not learnt new information at all. This process is usually approximated numerous times as a Monte-Carlo simulation.

We believe approximating the way people share information by only giving them one chance to do so is not entirely accurate; just because someone does not immediately share some news they hear with others,

it does not mean they never will. Because of this, our model modifies the independent cascade model so that when people gain new information, they have a certain probability of sharing the new information that gradually decreases over time, rather than immediately dropping to zero after one time interval.

2.2 Poisson Processes

A Poisson process is a continuous process which is made up of a collection of "arrivals", where the number of arrivals over the course of a given time period is given by a Poisson random variable. In addition, there is an arrival rate λ. Arrivals happen in continuous time and, assuming λ is constant, the time between arrivals is distributed as an exponential random variable with a rate parameter λ.

In this model, an arrival defines when a group of people at a certain vertex in the graph who don't know or believe some information yet are told that information at a rate λ until they know or believe it. λ is defined as the average rate of information spread per unit of time for a given location.

2.3 Monte-Carlo Method

Computationally, tracking the spread of information in a world of seven billion people in a model is unrealistic. In order to solve this problem, therefore, we use Monte-Carlo simulations (repeated sampling of a process) with a smaller amount of nodes to approximate information propagation.

3 Historical Background

Throughout the five periods of U.S. history we consider for our model, there have been various limitations to accessibility for each information medium. This includes limitations to how quickly information could spread via each medium. In the following section, we present a historical background for each information medium, including statistical data.

3.1 Newspapers and Telegrams

In the year 1800, there were only 200 newspapers being published in the United States, but by 1860, the number rose to 3,000. [**?**] By 1900, when the US population grew to 76,212,168 [**?**], there were two newspapers in circulation per individual. Although newspapers had massive popularity and circulation, they suffered from some drawbacks that hindered the reach of their information:

- **Newspapers had to be printed.** Unlike more modern forms of news where an individual can quickly write a blog post or make an announcement on TV, newspapers had to be printed and delivered to people for the news to be seen. Most papers circulating were printed either daily or weekly [**?**], so, at the fastest, people would be updated on yesterday's news today.
- **The speed at which newspapers' information could be delivered.** Although telegrams could deliver stories across the country in about four minutes, sending telegraphic messages was costly and was saved mostly for sending the biggest stories. [**?**] This meant that most newspapers were limited to local stories unless an important national event occurred.
- **Illiterate people could not gain information from newspapers.** While white Americans had an illiteracy rate of only 20% by 1870, black Americans had an illiteracy rate of 80%. [**?**] This means that of the 33,589,377 white people and 4,880,009 black people living in the U.S. [**?**], a total of approximately 10,600,000 were illiterate and could not directly gain information from a newspaper, so only about 75% of the U.S. population could get information from newspapers.

We assume the only drawback of newspapers by 1920 was the literacy rate, which had already risen to about 90% for black and white Americans combined. Today, the U.S. literacy rate is over 99%. [?]

3.2 Radios

Radio broadcasting became publicly available in the U.S. in 1920. By 1930, a total of 60% of American families owned a radio. [?] This period started a national American pastime of families listening to radio stories together every evening. Additionally, NBC and CBS were founded in 1926 and 1927, respectively, and both began broadcasting the same stories nationally. For the first time in U.S. history, people from different coasts could hear the same news at the same time. [?]

Although early radio news involved broadcasters simply reading newspaper stories on air, soon broadcasting companies began developing their own news stories. [?] While radio in the 1920s did not have as many drawbacks as newspapers did, there were still a few:

- **People had to be in one location to hear radio news.** While radios closed the distance between national and local news stories, those news stories could only be heard in specific locations, usually people's homes. This meant that people could not access the information radios broadcast at all times.
- **Radio news shows were broadcast in the evenings.** This means that, just like newspapers, radio news stories would generally be broadcast in 24 hour cycles, [?] although this has increased in recent years.

3.3 Television

Only a couple thousand American homes had televisions in the late 1940s. By 1955, however, over half of U.S. homes had one and by the end of the 1950s, TV was the most popular news medium in the U.S. [?] By the 1970s, 95% of Americans reported TVs as their main source of news. [?] Television had similar drawbacks to radios, however:

- **People had to be in one location to see TV news.** Much like radio news, Americans generally had TVs in their homes, which meant they could only gain TV information while being in their homes with the TV on.
- **TV stations broadcast news at specific intervals.** Much like radios again, TV stations would broadcast news stories at different intervals between commercials. This meant that people not only had to be in their homes watching TV to gain information from a TV, but they also had to be watching at a specific time. Unlike radios in the 1920s, however, television stations broadcast new information more frequently than once a day. [?]

3.4 Early Internet

While a computer network similar to the internet known as ARPAnet existed already in the 1970s, it was used only by computer science researchers. [?] The web as we know it today did not come around until CERN finalized its development and released it in 1991. [?] This early form of the web had very few users; even though fewer than 1% of the U.S. population had access to the web (about 2 million people), 73% of all people who had internet access globally were in the United States. [?] The percentage of Americans with internet access grew to 14% by 1995, however, and was over 50% by 2000. A total of 87% of the U.S. population had access to the internet as of 2015. [?]

Most of the drawbacks of the previous mediums of information were solved with the internet, especially how quick it made access to global communication for the first time in human history. The early internet had these limitations to its ability to spread information:

- **Low connection speeds.** Although the internet gave people more freedom in when they accessed information, early internet connections suffered from low speeds. [**?**] Today, LTE data and WiFi on phones can pull up websites in a matter of seconds, but even connecting to the internet in the 1990s took many minutes, which limited how quickly people could access information.
- **People had to be in a location with a computer to access internet.** Just like TV and radio, people could not carry the internet with them as they went places. Most Americans used the internet on computers in their homes, which meant people had to be at home access information on the internet.

3.5 Smartphones

Smartphones are arguably the fastest spreading technology in human history: from 2007 to 2010 smartphone ownership in the U.S. went from 5% to 40%, with 64% of U.S. adults owning a smartphone in 2015. [**?**, **?**] Smartphones have completely changed the way humans communicate with each other and access information. While all the previous mentioned mediums of information either traveled and updated slowly or required people to be in a single location to access information (or both), accessing global information with a smartphone can take seconds.

The rise of social media has also created a more connected world, allowing for a much faster spread of information than ever before. A total of 75% of surveyed U.S. smartphone users in October 2015 used their phone for accessing social media during the span of a single week. Because of how connected the world has become due to social media and smart phone access, the only limitation of smartphones' ability to spread information is the amount of people who own one (approximately 37% of the world population in 2015 [**?**]). With that being said, the rate at which smartphone ownership is growing means this will soon no longer be a drawback; it is estimated that 70% of the world's population will own a smartphone by 2020. [**?**]

A final limitation of information accessibility in present day is political regimes blocking citizens from accessing the internet. Despite how much information is available online today, citizens of North Korea, for example, cannot access most of it due to strict government regulations [**?**], which means news stories that spread quickly across the rest of the planet do not spread at the same rate in North Korea. Our model does not truly take this limitation into account, however, because, as mentioned in our assumptions, there are too many factors to consider for any country in such a situation.

4 Description of Model

4.1 Parameters

This paper aims to analyze the spread of a piece of information compared to that information's inherent value. One issue related to this is defining exactly what "inherent value" means since the term can be so subjective. For our model, we use several parameters to help us quantify some information's "inherent value" and how quickly this information will spread. These parameters are shown in the following table and defined below.

Variable	Description
t_i	Length of a news cycle
A	How compelling a piece of information is
k_m	Reach of a particular medium of information
p_t	Accessibility
I	Information persistence

- **Δt: Length of a news cycle**

  Δt a value inherent to the medium of information and is equal to the amount of time between each news cycle for every vertex. For example, if $\Delta t = 1$ hour, then every vertex releases news of that type once an hour. For an origin of these values, see the appendix.

Time Period	Δt
1870s (Newspapers)	24 hours
1920s (Radios)	24 hours
1970s (Television)	12 hours
1990s (Internet)	4 hours
2010s (Smartphones)	1 hour

- **A: How compelling a piece of information is**

  A is a parameter inherent to a specific story and defines how compelling a certain story is to a large amount of people. In particular, A is equal to the expected number of times a vertex will broadcast a story during a given news cycle. For example, news of a president's assassination will have a high A value (especially in the late 1800s since there were so many newspapers in circulation), while a wedding in a small town will have a low A value in both the 1800s and today. Note, however, that this parameter assumes the story is not forgotten at all and that distance between locations does not affect the rate at which news is spread. This value is derived from an assumption based on how many types of a specific medium were in circulation (e.g., there were a lot of newspapers but not as many TV channels).

- **k_m: Reach of the medium of information**

  k_m is a parameter inherent to the medium of the information (e.g. word of mouth, newspaper, internet, etc.). In particular, k_m represents how distance between cities affects the spread of new information. This only affects the probability that information jumps from one node to another, not the physical rate at which information travels. This parameter is assumed to be proportional to the radius of the city from which the news originates. In addition, k_m can vary depending on the importance of the message being sent. For example, in the 1870s, a story with a low A value would not be sent using a telegram because the cost would not be worth sending it, so k_m will grow considerably. However, a story that is important (high A) will be sent via telegram, so k_m will be low. For the origin of each value of k_m, see the appendix.

$$k_{newspaper} = \begin{cases} 0.67, & A \leq 100, \text{ newspaper only} \\ 0.4, & A \geq 1000, \text{ telegram} \end{cases}$$

$$k_{radio} = \begin{cases} 0.67 & A \leq 50, \text{ local radio only} \\ 0.4 & A \geq 300, \text{ after founding of national stations} \end{cases}$$

$$k_{tv} = 0.01 \text{ for all stories}$$

$$k_{internet} = 0.0001 \text{ for all stories}$$

$$k_{smartphones} = 0.00001 \text{ for all stories}$$

- $p_{m,t}$**: Accessibility parameter**

  The parameter $p_{m,t}$ is also inherent to the information medium and defines the accessibility of that source. In particular, $p_{m,t}$ is the proportion of the U.S. public with access to medium m at time t. For example, this parameter is zero for the internet in the 1870s (internet had yet to be invented). All derivations of this parameter are from the "Historical Background" section above.

Time Period	$p_{paper,t}$	$p_{radio,t}$	$p_{tv,t}$	$p_{internet,t}$	$p_{smartphones,t}$
1870S-1890s	0.75	0	0	0	0
1920S-1930s	0.90	0.60	0	0	0
1950S-1970s	0.98	0.90	0.95	0	0
1990s	0.99	0.90	0.98	0.14	0
2010s	0.99	0.90	0.97	0.87	0.64

- I**: Information persistence**

  The parameter I is equal to the amount of time from when a particular location discovers information until the rate of sharing that story has been reduced by $1/e$. This could be due to the fact that people find the story less interesting, people have forgotten about the story, etc.

4.2 Graph Explanation

While our model is scaled down for optimal computational complexity, the behavior of our graph mimics the behavior of a real informational network. Each vertex in the graph represents a small collection of people. These vertices are assigned randomly, but in proportion to cities' populations. They are placed spatially according to a radially symmetric, two-dimensional normal distribution. In particular, the center of the normal distribution is the city center in terms of real latitude and longitude values and the standard deviation is equal to the approximate radius from the center of the city to the city limits. Each vertex is randomly connected to an average of 6 other vertices via a normal distribution with a standard deviation of 2.

In the computational part of our model, we constructed a standard, bidirected graph with vertices distributed according to the top 1000 most populous cities in the United States (as of modern day; however, we assume the most populous cities today have been the most populous cities throughout history). Each vertex is placed randomly with the probability of it being placed in a certain city proportional to the population of that city. In addition, each vertex is placed according to a normal distribution with a standard deviation of one degree latitude and one degree longitude.

Each vertex has the following properties:

- **Informed Status:** a Boolean value of true or false indicating whether the group knows a piece of information or not.
- **Bias Value:** a bias value dictating the group's bias towards any particular story. This value is randomly assigned according to a normal distribution with mean zero and variance one.

4.3 Formula

Aside from the above mentioned parameters, our model also uses the variables defined in the following table as part of its formula:

Variable	Description
t_i	the time passed since vertex i gained information
d_i	physical distance between node i and its neighbor
M_t	set containing all mediums for a particular time period
N_i	set containing all informed vertices that enter vertex i
T	period of time the network is modeling

If a certain vertex i receives information, there is a positive probability this node will spread that information to one of its neighbors. This process is modeled as a Poisson process with rate parameter α_i, where:

$$\alpha_i = \sum_{m \in M_t} p_{m,t} A e^{-(t_i/I + k_m d_i)} \tag{1}$$

Note that α_i is a sum since vertex i is broadcasting information via all possible news sources available in that time period, even ones that are no longer the main source of information.

The overall rate at which a story "arrives" to any vertex i in this Poisson process, then, is:

$$\lambda_i = \sum_{j \in N_i} \alpha_j \tag{2}$$

This formula results in the following diagram of how information propagation occurs in our model:

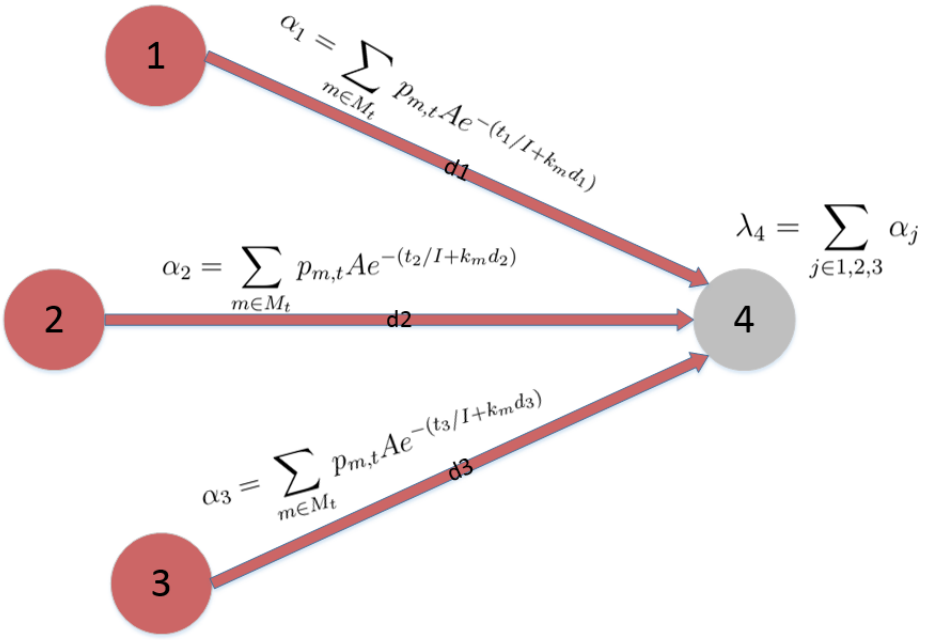

Figure 1: Node 4 is receiving information with a rate of λ_4

5 Modeling Information Networks

5.1 Model Results

We ran our model across all five time periods (1870s, 1920s, 1970s, 1990s, and 2010s) with four types of stories, where interesting corresponds to a high A value and memorable corresponds to a high I value:

- **interesting and memorable**, e.g., a president's assassination.
- **interesting but not memorable**, e.g., an adorable cat video or a pop star's wedding.
- **not interesting but memorable**, e.g., a yearly economic report.
- **not interesting and not memorable**, e.g., someone posted on Facebook that they ran a 5K.

In doing this, we created a plot of the proportion of vertices who became informed as a function of time and a plot showing all vertices geographically in terms of latitude on the vertical axis and longitude on the horizontal axis, color coded by whether or not they had been informed of the news (blue for uninformed, red for informed).

Figures for modeling the spread of various information sourced in New York City are shown below.

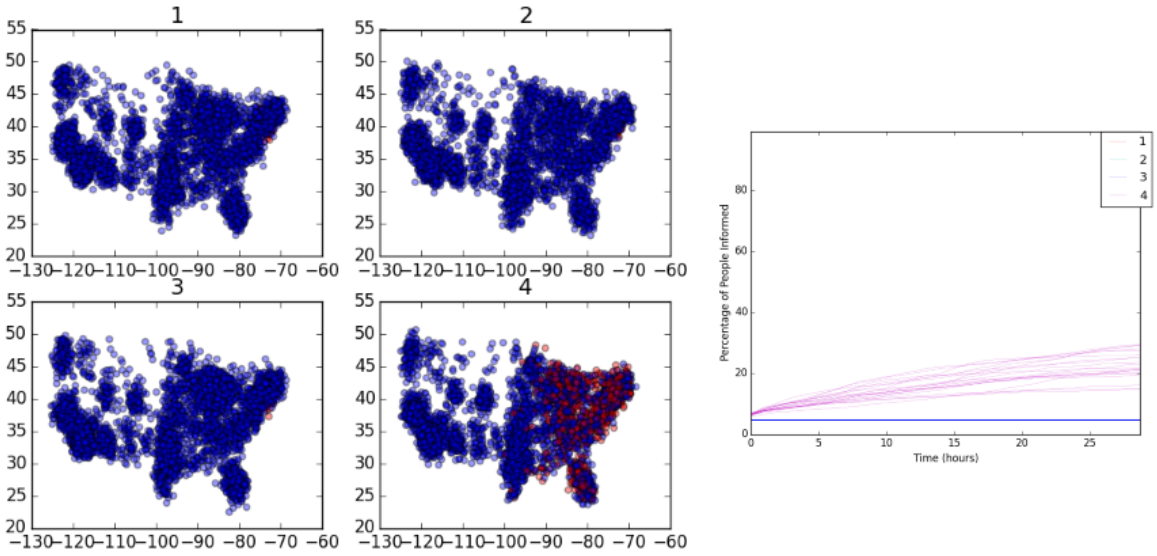

Figure 2: **1870's** (A, I) values: 1 - (50, 100), 2 - (100, 500), 3 - (50, 100), 4 - (1000, 500)

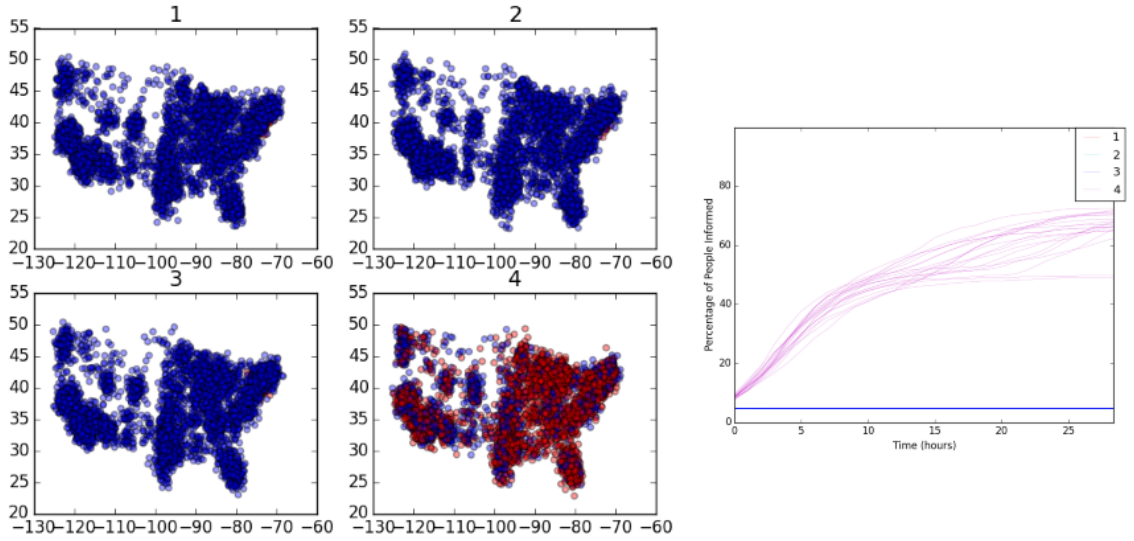

Figure 3: **1920's** (A, I) values: 1 - (25, 100), 2 - (50, 500), 3 - (25,100), 4 - (300, 500)

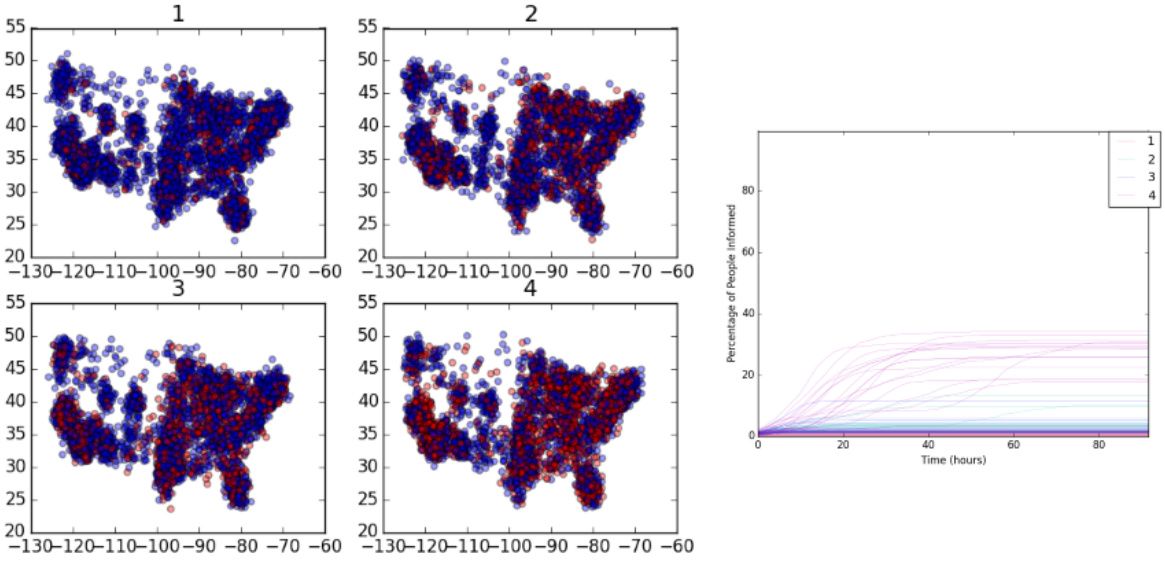

Figure 4: **1970's** (A, I) values: 1 - (10, 15), 2 - (10, 50), 3 - (40, 15), 4 - (40, 50)

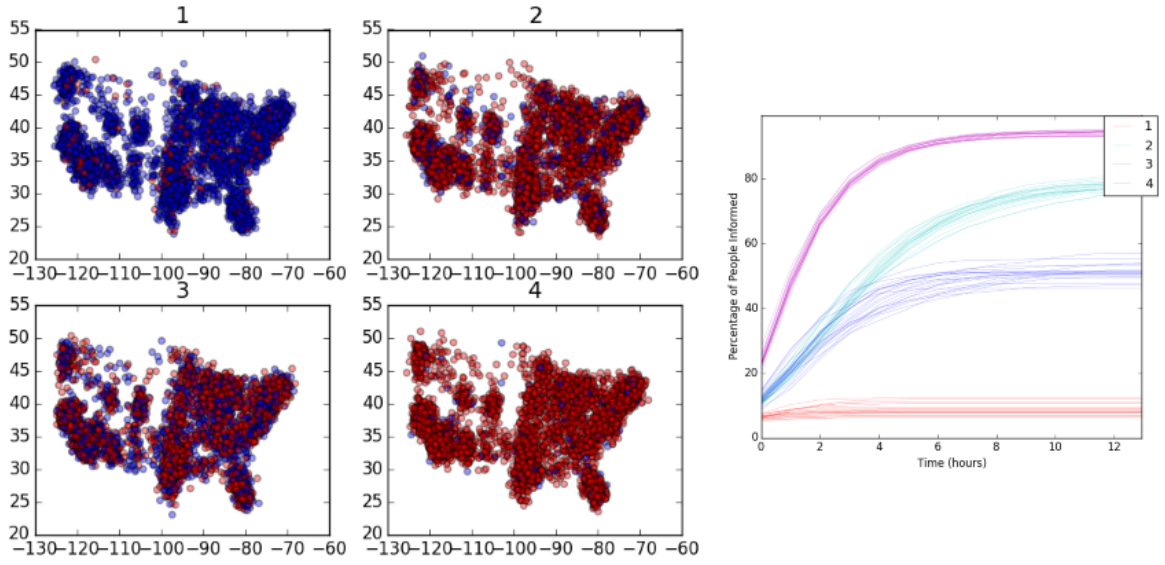

Figure 5: **1990's** (A, I) values: 1 - (2, 2), 2 - (2, 7), 3 - (5, 2), 4 - (5, 7)

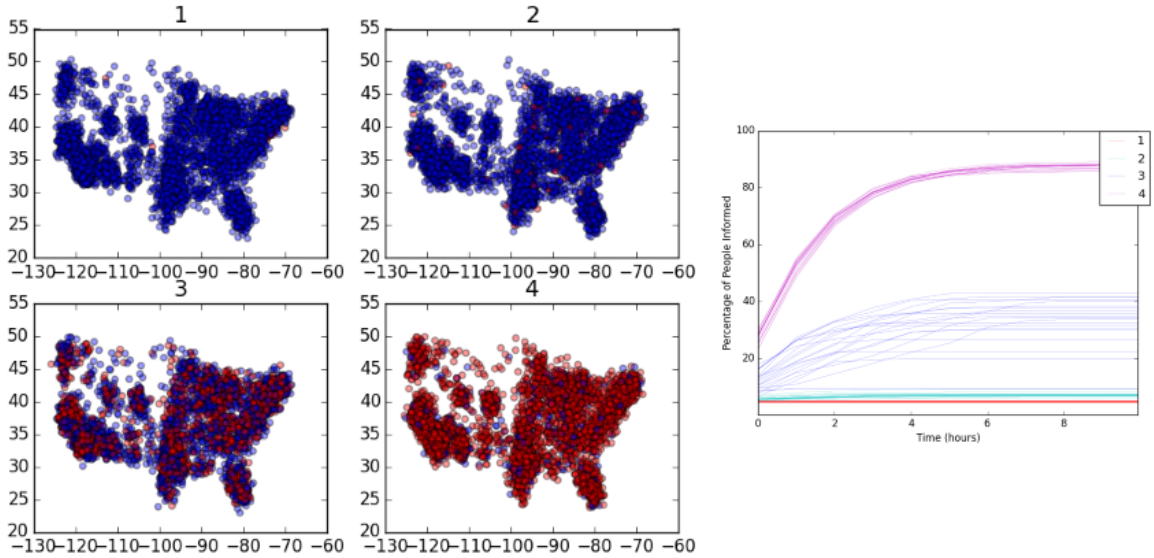

Figure 6: **2010s** (A, I) values: 1 - (0.1, 0.3), 2 - (0.1, 1), 3 - (1, 0.3), 4 -(1, 1)

We can see how in the 2010s or 1990s, stories do not need very high A and/or I parameters to go viral and spread all across the U.S. In comparison, in the 1870s, only stories with very high A and I values spread across the U.S. due to the financial and time cost of transmitting information across the country. Information with high A and I values in the 1920s also takes off on a national scale after a while due to national broadcasting stations standardizing stories across the country.

5.2 Applying Our Model to 2050

Using our model and modifying its parameters based on extrapolations from historical data, we can predict how information will spread across the U.S. in 2050. We change one paramter specifically to do this: $p_{m,t}$, the probability a person has access to a medium m during period t.

As of 2015, 85% of young adults (ages 18-29) in the U.S. own smartphones and only 27% of Americans over 60 do. [**?**] We assume that by 2050 the majority of senior citizens in America will have passed away and the current generation of young adults will be considered adults, meaning at least 85% of Americans will own a smartphone. We assume this number to be at least 99% though, given that if the majority of young adults currently own smartphones, the teenagers and children of 2015 and the children of 2015's young adults will have smartphones 35 years from now as well. Therefore, the changed parameter for our model of the U.S. in 2050 is $p_{smartphones,2050} \approx 0.99$.

As cited earlier, we also know current estimates predict 70% of the world population will own a smartphone by 2020. [**?**] Only 5% of U.S. citizens (about 20 million people) owned a smart phone in 2007 after the iPhone's release in America— arguably the biggest catalyst in increasing global smartphone ownership. [**?**] This means the number of smartphone owners globally will have gone from 20 million to approximately 6 billion in just 14 years. Because of this, we assume the global smartphone ownership rate will rise to at least 99% by 2050 as well.

Since k_m and Δt are values inherent to the news source and do not change with time, we leave those constant based on the assumptions that news can already travel globally with an incredibly quick pace. With all of this kept in mind, we compare the new accessibility parameter of 2050 to 2010, with all other parameters kept equal over time in the figure below.

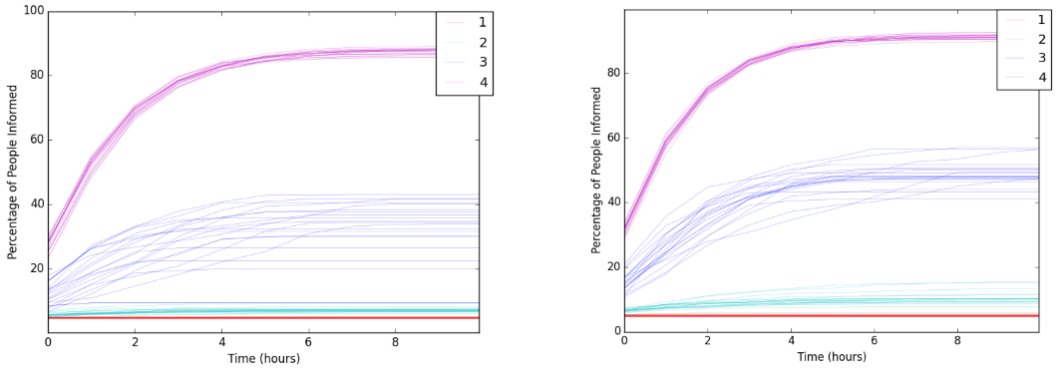

Figure 7: **2010s vs 2050s** (A, I) values: 1 - (0.1, 0.3), 2 - (0.1, 1), 3 - (1, 0.3), 4 - (1, 1)

We can see that the most interesting stories will still reach the same number of people, but some stories that are simply memorable (economic reports) or compelling (cat videos) will become more circulated. This shows that the rate of information propagation has mostly reached its peak in present day, but the amount of information saturation in peoples' lives will continue increasing over the coming years.

6 Analysis of Model

6.1 Defining "News"

In our model, we consider "news" to be any information that spreads from the city it originated. In order to derive this mathematically, we look more carefully at the mechanics of our model and the properties of a Poisson process.

In a Poisson process, the expected number of arrivals over all time can be calculated. Since news is spread in our model according to a Poisson process, the expected amount of news spread by one vertex can be calculated according to the following:

In a Poisson process with rate parameter λ, the expected number of arrivals ($E(N)$) in a time step Δt is:

$$E(N(\Delta t)) = \lambda t$$

So the expected number of arrivals in a time step dt is:

$$E(N(dt)) = \lambda dt$$

Therefore, the expected total number of arrivals is:

$$E(N) = \int_0^\infty \lambda(t) dt$$

This means that if vertex i directs to X_i of its neighbors, the expected number of neighbors vertex i informs over all time is given by:

$$E(N) = \int_0^\infty \sum_{j \in X_i} \alpha_j dt$$

Now assume vertex i has at least one uninformed neighbor. If the vertex has no uninformed neighbors, we can say the information has fully spread from vertex i, and we can focus our attention on a different vertex. With this is mind, we can extract the term in the sum containing the closest neighbor to vertex i, so:

$$E(N) \geq \int_0^\infty p_i A e^{-(t_j/I + k d_j)} dt$$

$$E(N) \geq p_i A e^{-k d_j}$$

Where d_j is the distance to the closest neighbor of vertex i. If this expected value is greater than one, then the news will spread without an upper bound *as long as each vertex has a neighbor close enough to it*. To figure out exactly what values of d_j enables this growth, we solve the following for d_j:

$$p_i A e^{-k_n d_j} > 1$$
$$\frac{ln(p_i) + ln(A) + ln(I)}{k_n} > d_j$$

If the distance to the closest neighbor of any given vertex in the graph is less than the value above, the information will spread; therefore, spreading information qualifies as "news" in any area (since d_j depends upon location) where these conditions are satisfied.

6.2 Validation of the Model

In order to test the accuracy of our model, we compare how information propagates in our model to a real event using Google Trends. In particular, we look at the number of web hits of *Kony 2012*, cited by TIME as the most viral video of all time. [**?**]

Kony 2012 is a short film by Invisible Children that depicts the atrocities of Joseph Kony, a Ugandan war leader and international war criminal, including specifics such as his recruitment of child soldiers. Although initially gaining huge traction, the film was met with controversy citing its validity and shortly faded from relevant news. [**?**] We approximated the parameters in our model that would represent the spreading of a viral video.

The full comparison can be seen below:

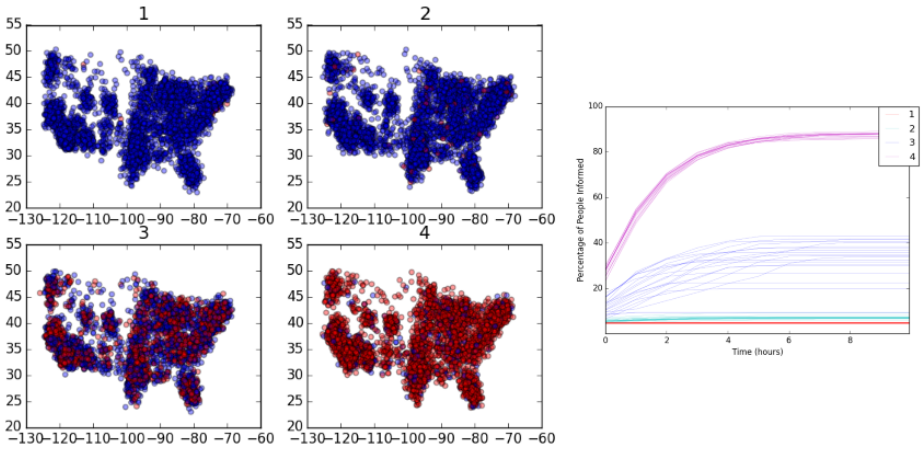

Figure 8: **2010s** (A, I) values: 1 - (0.1, 0.3), 2 - (0.1, 1), 3 - (1, 0.3), 4 - (1, 1)

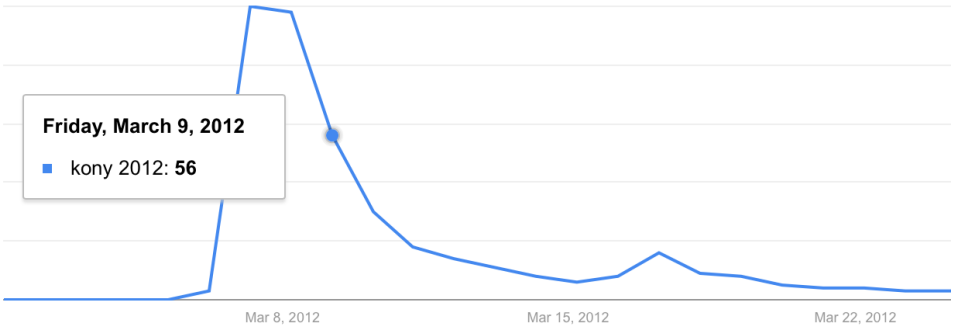

Figure 9: Trending Line of *Kony 2012* in March of 2012. [?]

An initial problem to note in comparing these two graphs is that the y-axes do not represent the same information. While our model shows the percentage of people who know about a topic, the Google Trends plot shows the number searches for a topic relative to the total number of searches done on Google at the time. We consider it safe to assume, however, that based on the high proportion of Americans with smartphones and internet access in 2012, most people would be informing themselves of any trending topic using Google search, so we can accurately compare this data to our model.

While the Google Trends plot has a larger time step, the jump from zero to maximum saturation still takes place over a single time step, just like in our model. As a result, our model's estimate of approximately ten hours for a highly compelling story to reach high saturation is consistent with the real trend of the *Kony 2012* video.

The Google Trends line also shows an approximately exponential decay in terms of interest in the video. We can see this because the video dropped from its max of 100 relative searches on March 8, 2012 to just 56 on March 9. This supports our model as well because our I parameter assumes people lose interest in a story at an exponential rate. The actual rate at which interest dies down in the Google Trends line is much slower than our model, however.

Although we validated our model only with a trending story from modern day, we assume the nature of trending information has not changed much throughout time, i.e., if a story is interesting enough, people will stay uninformed of it and then become informed of it in groups at an exponential rate, despite whatever medium spreads that information. We believe the only factor mediums of information have affected since the 1870s in the case of interesting or viral stories is the time it takes for a story to take off, given that some mediums have had longer travel times across distances.

7 Information Influence

7.1 Theories and Concepts of Information Influence

Informational social influence is a phenomenon that occurs whenever a person's emotions, opinions, or behaviors are affected by others [?]. It is possible to examine the *Kony 2012* example from section 6.2 as spreading by social influence, rather than simply because it is interesting information. People on social media were compelled to care strongly about the issue of Joesph Kony because it appeared to be a consensus among their peers that he needed to be stopped, even if actual Ugandans criticized the video. [?]

While our model has been used to track the spread of information against a defined, qualitative inherent value, the theory behind social influence can be used to further generalize our model. In particular, since our model involves vertices interacting with and influencing one another, we can determine how different parameters can effectively influence public opinion and interest.

7.2 Utilizing Information Influence

- **Topology and strength of the network.** The graph's topology strongly affects our model's behavior since we are modeling information propagation within a network. Since we randomly distribute vertices around known locations of cities, we expect those locations to form clusters of nodes. These clusters play major role in the 1870s specifically, due to the difficulty in transmitting stories nationally.

  Considering the network's strength, it is clear that adding more edges to a graph increases the rate at which information spreads. In our model, this is specifically due to the fact that the rate of arrival of information into a vertex is equal to the sum of the broadcasting rates of its neighbors. If the number of neighbors goes up for a particular vertex, the rate at which information is received by it goes up as well. In addition, the more connected the world is, the more likely there is a connection from obscure places to a vertex close to the original source, so the overall spread of the information goes farther. All of this can be seen in the figure below.

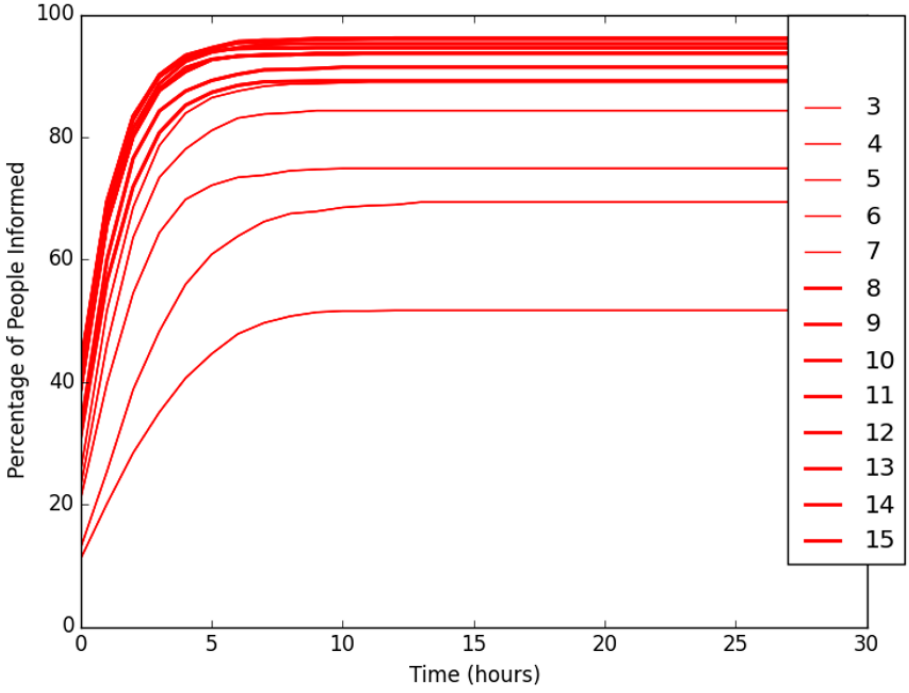

Figure 10: Rate of Information Spread versus Number of Edges per Vertex

- **Bias.** Another property that news sources can take advantage of in order to spread a product or an idea is human bias. In order to quantify this, we introduce a new bias parameter, B, that represents how much more likely a vertex is to share information based upon its own bias (we will call this B_i) and the bias of the information it is spreading (we will call this B_s). In particular,

$$B = \frac{|B_i| + 1}{|B_i - B_s| + 1}$$

$$\alpha_i = BAe^{-(t_i/I + kd_i)}$$

where B_i is assigned for a vertex via a normal distribution with a standard deviation of one. A biased vertex (determined by the value of B_i) is more likely to share information in general (looking at the numerator), but a vertex is less likely to share information that is biased in a way that does not line up with its own bias. For example, a vertex that is not biased ($B_i = 0$) will not be more likely to share an unbiased source, but less likely to share sources that are very biased. In addition, biased vertices ($|B_i| > 0$) will be more likely to share sources that match their own bias, equally likely to share unbiased sources compared to unbiased vertices, and less likely to share sources that are biased in the opposite direction.

However, bias only affects the rate at which a node broadcasts information, not the rate at which it is willing to receive it. Whether or not a biased person believes in a story, the story will reach them. This highlights a weakness in our model: it works to discover how many people *know about* a story, but not how many people *agree with* that story, although the two are correlated.

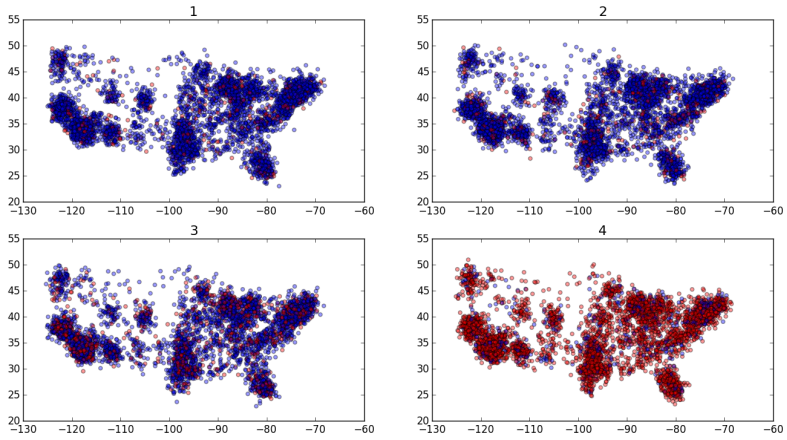

Figure 11: Graph with Negative Bias in California. $|B_i| = -1, |B_s| = 1$. (A, I) values: 1 - (0.1, 0.3), 2 - (0.1, 1), 3 - (1, 0.3), 4 - (1, 1)

- **Information source.** The source of a story can greatly affect how quickly it spreads. For example, the following figures show results when the same information starts in Los Angles versus New York City, which has been the source of all information throughout the paper to this point.

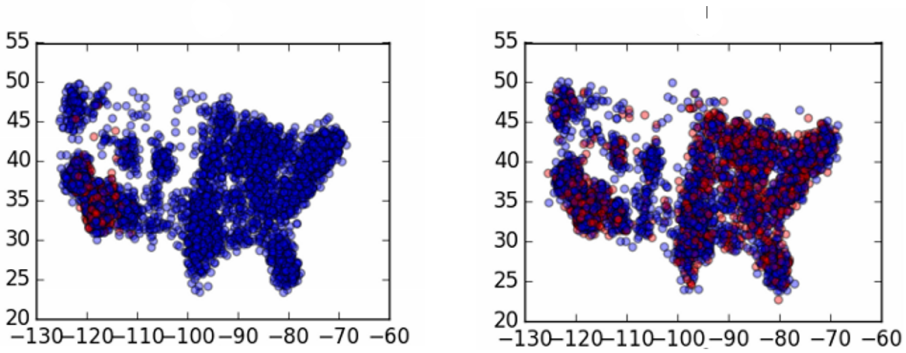

Figure 12: Info starting in LA versus New York in 1970. ($A = 10$, $I = 15$)

Since New York is surrounded by more vertices than LA, information spreads farther, even though all other parameters are the same.

8 Strengths and Weaknesses of Our Model

8.1 Strengths

- **Flexible parameter for the rate information spreads.** Our model utilizes the rate at which information randomly travels between vertices on a graph. The parameter that keeps track of this rate between any two nodes in particular is α_i. This parameter has several intuitive terms that describe how it grows or decays over time and distance. α_i can easily be modified depending on the type of information or the medium of the information. For example, if a story is expected to lose interest when it is farther away from its source, a term depending on the distance from the *original* source of the story (rather than the distance between two adjacent vertices) can easily be added into the formula for α_i.
- **Combines continuous and binary methods.** While our model is similar to the independent cascade model, the latter assumes that there is only one chance for a vertex in the graph to inform a neighboring vertex, essentially making it a binary model (either a neighbor is informed or not). By incorporating characteristics of Poisson processes, our model not only takes into account that people share information with a certain probability after being informed, but also that people are more likely to be informed by a piece of information if a lot of people connected to them know that same piece of information.
- **Parameter values based upon real data.** All parameters used in our model are derived from quantitative historical data. Although we made assumptions for exact numbers due to the lack of accurate historical data, we derived the numbers that describe access to information mediums, area of influence of informational mediums, and the length of a news cycle from historical data.

8.2 Weaknesses

- **Lack of a time delay.** There are several assumptions that go into our model that oversimplify the complexities of the reality of spreading information. One of these assumptions is that all distance reduces the probability that information will be spread at any one particular moment. Information is assumed to travel from one place to another instantaneously once it has been determined that it

will travel, but this may not necessarily be the case for the 1800s and 1920s when it took longer for information to travel.

- **Independence of parameters.** All attributes of the vertices in our model are attributed randomly (i.e location, bias, and number of connections) and are assumed to be independent. This ignores the reality that, among other things, political affiliation tends to be very strongly correlated with geographical location.
- **Our model measures knowledge, not agreement.** When running our model determine the effect of bias on how people agree with a biased source, little change was observed in the distribution of knowledge. This is because the bias parameter only affects how frequently a given node will output information, which is offset by other people who are broadcasting the same story. Even though the model concludes that biased information spreads approximately the same distance as non-biased information, the model can not conclude what proportion of people agree with the biased news or how they rate its inherent value.
- **Arbitrary step in A for k_m.** The values for k_m, which model the area of influence of a specific information medium, change based upon the importance of the story. Since it is difficult to determine exactly when a story can be considered "compelling" enough for national news, the values of A at which the k_m values change for a particular time period are somewhat arbitrary. In a future version of the model, more research should go into making A more concrete.

9 Future Improvements and Conclusions

The problem of accurately modeling a system as unpredictable as a large-scale information network is a difficult one.

Our model is a flexible first-order approximation for the spread of information, influence, and trends over several time periods in U.S. history. This gives mathematical intuition behind the physical phenomena of information flow. Given the parameters we chose to put into our model, we found historical statistics to try matching these to the real world as closely as possible. However, several factors were not taken into account. These include how the time of day a story originates (especially when considering global time zones) could affect information spread, political factors that prevent people in certain locations from accessing news, etc. We also made many assumptions to simplify how our model works that may have diminished its accuracy.

Analyzing various periods of U.S. history shows that the ability of a news medium to travel can be more important than news itself, especially when it is difficult for a medium to travel long distance quickly. Our model also shows that people tend to acquire new and exciting information in large groups at a quick exponential rate and then forget about that information in a similar fashion. In addition, extrapolating our model to 2050 predicts that more information will spread at a faster rate in the future and the average inherent value of information will decline due to information over-saturation.

A Appendix

A.1 Selection of k_m Parameters

Each k_m was determined by estimating the radius of the area of influence for each news medium in terms of the average radius of a city (in this model, the radius of a city is 1 degree latitude - or about 111 km). The inverse of this value is approximately the distance needed for the sphere of influence to drop by a factor of $1/e$.

Newspaper area influence was determined by the number of readers of *The Denver Post* [**?**] divided by the number of newspaper readers in the Denver metro area (p_{2010} for newspapers multiplied by the population of Denver [**?**]). The square root of this value is equal to the radius of influence of local news.

Local radio influence as well as local television influence was determined using a plot from the Federal Communications Commission. [**?**]

Telegraph and national radio/television area of influence was assumed to be national, so the diameter of the United states was used as the radius of influence of these mediums.

The area of influence of the internet was assumed to be global for all stories.

A.2 Selection of Δt Values

Values for Δt are the estimated length of one news cycle for each medium.

Newspapers in the late 1800s were printed and delivered daily (at the fastest; some were printed and delivered weekly still). [**?**]

Most radio news stories were broadcast on average about once an evening in the 1920s and 1930s. [**?**]

For television, due to a lack of accurate historical data, we estimated using statistics from modern day. Since round 33% of people report getting news at all times of the day[**?**], and the rest of people report getting news at one specific time, we approximated this as twice a day in the model.

For the internet before social media, this number is estimated using approximate times and effort required to upload content to the early internet

Social media times are estimated from numbers which include 45% [**?**] of Facebook users alone who visit the website multiple times a day, as well as the capabilities of notifications on the platform of a smart phone for important stories.

B References

[1] Jong-Ryul Lee, Chin-Wan Chung. *Scalable influence maximization for independent cascade model in large- scale social networks.* 2012.

[2] University of Illinois at Urbana-Champaign Library. *American Newspapers.*

[3] U.S. Census Bureau. *June 1, 1900 Census.*

[4] U.S. Census Bureau. *June 1, 1870 Census.*

[5] Mitchell Stephens. *History of Newspapers.* https://www.nyu.edu/classes/stephens

[6] R. Bond *The Handbook of the Telegraph.* 1870.

[7] Max Roser. *Literacy.* 2015. http://ourworldindata.org/data/education-knowledge/literacy/

[8] *The growth of radio in the 1920's.* 2011. http://www.mortaljourney.com/2011/04/1920-trends/radio-history

[9] Mitchell Stephens. *History of Television.* https://www.nyu.edu/classes/stephens

[10] Barry M. Leiner, Vinton G. Cerf, David D. Clark, et al.. *Brief History of the Internet.* http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet

[11] World Wide Web Foundation. *History of the Web.* http://webfoundation.org/about/vision/history-of-the-web/

[12] World Mapper. *Internet Users 1990* http://www.worldmapper.org/display.php?selected=335

[13] World Bank. *Internet users (per 100 people).* http://data.worldbank.org/indicator/IT.NET.USER.P2?page=3

[14] MIT Technology Review. *Are Smart Phones Spreading Faster Than Any Technology in Human History?* 2012. http://www.technologyreview.com/news/427787/are-smart-phones-spreading-faster-than-any-technology-in-human-history/

[15] Aaron Smith. *U.S. Smartphone Use in 2015.* 2015. http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/

[16] Ingrid Lunden. *6.1B Smartphone Users Globally By 2020.* 2015. http://techcrunch.com/2015/06/02/6-1b-smartphone-users-globally-by-2020-overtaking-basic-fixed-phone-subscriptions/

[17] U.S. Census Bureau. *World Population Estimates: 1950-2050.* 2015. https://www.census.gov/population/international/data/idb/worldpopgraph.php

[18] The Denver Post. *Denver Post Sunday Daily Circulation.* 2013. http://www.denverpost.com/ci_23138277/denver-post-sunday-daily-circulation-climb

[19] U.S. Census Bureau. *2014 Census.*

[20] Federal Communications Commission. *Denver radio and television station broadcast range.* 2008. https://transition.fcc.gov/dtv/markets/maps_current/Denver_CO.pdf

[21] Nielsen. *The Cross Platform Report: A Look Across Media.* 2013.

[22] Communic@tions Management Inc. *Sixty Years of Daily Newspaper Circulation Trends.* 2011. http://media-cmi.com/downloads/Sixty_Years_Daily_Newspaper_Circulation_Trends_050611.pdf

[23] American Press Institute. *How Americans Get News.* 2014.

[24] Maeve Duggan, Nicole B. Ellison, Cliffe Lampe, et. al. *Frequency of Social Media Use.* 2015. http://www.pewinternet.org/2015/01/09/frequency-of-social-media-use-2/

[25] Matthew Sparkes. *Internet in North Korea.* 2014. http://www.telegraph.co.uk/technology/11309882/Internet-in-North-Korea-everything-you-need-to-know.html

[26] Nick Carbone. *Top 10 Viral Videos of 2012.* 2012. http://entertainment.time.com/2012/12/04/top-10-arts-lists/slide/kony-2012/

[27] *Informational Social Influence.* http://changingminds.org/explanations/theories/informational_social_influence.htm

[28] Google. *Google Trends.* https://www.google.com/trends/explore

[29] Okwonga, Musa. *Stop Kony, yes. But don't stop asking questions.* 2012. https://originalpeople.org/stop-kony-yes-but-dont-stop-asking-questions/