Physics 2620J Final Project

Andrew Friberg
Brown University
Department of Physics
(Thanks to Clovis Wong for discussing these ideas with me)
(Dated: April 26, 2021)

A model of online activity is introduced, and the effects of belief updating functions, initial biases, and social media recommendations are considered. Comparisons are drawn to recent results published in *Nature*.

I. INTRODUCTION

Profit-driven media corporations have always catered to people's emotions and whims, and sensationalist news has existed since Hearst and Pulitzer first competed for newspaper sales. Vindictive and one-sided, such news dissemination can be extremely damaging and polarizing, to the point of starting and maintaining military conflicts. The weaponization of news was first employed en masse by the British army during the first world war, creating iconic posters and art that have had an impact on the world long after their original purpose was fulfilled. Figure 1 shows the famous "Earl Kitchener wants YOU" poster, alongside the American imitation of it.[1]

FIG. 1. An early example of the weaponization of information

The Americans were not the only admirers of British propaganda, and it eventually came back to haunt them in a big way. One German Prisoner of War was a huge fan of the simple depictions of the truth, which fortified the British fighting spirit and left no room for doubt. As Tim Wu recounts it, "The fan was Adolf Hitler, and given his chance, he thought he could do even better." [2]

In the age of the newspaper and telegram, news propagation was slow and done in broad strokes. In an age of plentiful data, personalized news, and instantaneous communication, polarization is usually not a result of government action, but more often the results of endlessly complicated algorithms that encourage the formation of thought echo chambers.

To try and quantify some of the questions of internet polarization, this report will summarize the results of a paper published in *Nature* by Sikder, Smith, Vivo, and Livan[3], and will also introduce a model of online interactions that tries to take into account certain complicating factors that were not present in Sikder's paper.

II. AN OVERVIEW OF SIKDER, SMITH, VIVO, & LIVAN'S PAPER

Sikder, Smith, Vivo, and Livan (The investigators) considered a social network consisting of n agents, each labelled by an index $i \in \{1, 2, ..., n\}$. Social connections were modelled via a graph, and each agents was deciding whether or not to believe in a hypothesis X based on information propagated through edges in the graph, with each agent receiving an initial stimulus $s_i \in \{+1, -1\}$. Time move forward in discrete intervals, and at each interval t, each agent shared their accrued vector of information $s_i(t)$. For instance, an agent i with neighbors j and k will have information vectors

$$\begin{aligned} \mathrm{s}_i(0) &= \{s_i\} \\ s_i(1) &= \{s_i, s_j, s_k\} \\ s_i(2) &= \{s_i, s_i, s_i, s_j, s_j, s_k, s_k, s_{d=2}\} \end{aligned}$$

where $s_{d=2}$ is the set of all stimuli coming from j and k's neighbors (stimuli from a distance of 2 away from i). The total number of positive stimuli at time t is $N_i^+(t)$ and the number of negative stimuli is $N_i^-(t)$. The researchers also defined the signal mix as:

$$x_i(t) = \frac{N_i^+(t)}{N_i^+(t) + N_i^-(t)} \tag{1}$$

The agents updated their belief using a modified naïve Baye's rule:

$$\begin{aligned} P(X|s_i(t)) &= \frac{P(s_i(t)|X) \times P(X)}{P(s_i(t)} \\ &\approx \frac{\prod_c P(s_i^c(t)|X) \times P(X)}{P(s_i(t)} \end{aligned} \tag{2}$$

which fails to consider the joint probability between dif-

ferent events, a standard error that often occurs in human decision making. The best estimate of X that the agent can make at time t is given by $\tilde{X}_i(t) = H(x_i(t) - 0.5)$ where $H(x)$ is the Heaviside step function.

Some nodes of the graph were biased agents, meaning that they ignored signals that conflicted with their beliefs, and replaced them with signals that concurred with their current belief $\tilde{X}_i(t)$. Biased agents ignored conflicting signals with fixed probability q, and the proportion of biased agents in the graph is denoted f.

Using analytical results derived from graph theory, the investigators derived analytic results for the steady-state mix of signal that unbiased agents received, as shown in Figure 2. The analytic results are very close to simulated results, especially for the k-regular graph on which the analytic results were derived.

A fascinating result of this study is that a set of agents with very little confirmation bias (q) is actually very susceptible to completely shifting to one viewpoint or another. For $\frac{1}{2} < q \leq 1$, biased agents "stay strong" in their beliefs even when they are in the minority, preserving a signal mix by filtering out opposing signals. But for $q < \frac{1}{2}$, even biased agents can be convinced to switch their views, if there are enough agents who have opposite views. Surprisingly, maintaining groups of agents who are intransigent in their views actually *maintains* a mix of discourse.

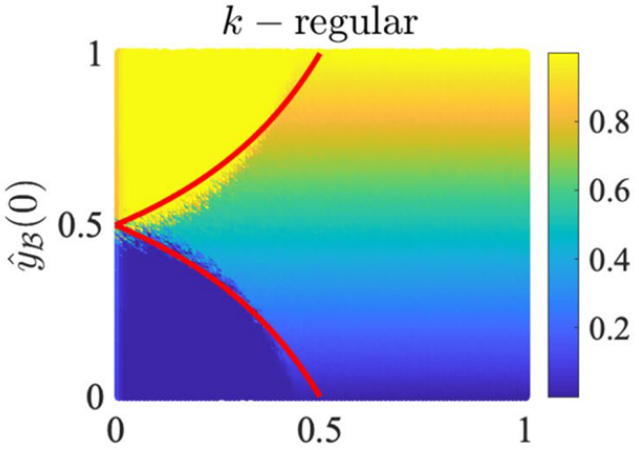

FIG. 2. The average steady state signal mix for unbiased agents (represented by the color gradient), as a function of confirmation bias q (horizontal axis), and fraction of biased agents who were positively oriented at $t = 0$, denoted $\hat{y}(0)$. Analytic prediction shown in red. Run on a k-regular graph. Image taken from Sikder et. al.

The last mathematically-approachable line of inquiry the paper considered was the *accuracy* of the network. This involved making a normative judgement; the researchers chose to indicate $X = +1$ as the ground truth, and defined the accuracy $\mathcal{A}(\mathcal{G})$ as the expected fraction of accurate agents in the steady state, accurate agents being those whose believe that $X = 1$. An important parameter is $p = P(s = +1|X = +1)$, the probability that a given signal will be positive given the base truth. High p indicates reliable news, low p indicates a prevalence of unreliable, or fake, news.

Curiously, or perhaps not so curiously given the previously discussed results about optimal confirmation bias to maintain mixed signals, they found that accuracy is not a monotonic function of f, the fraction of biased agents in the graph. Accuracy peaks at an optimal value of f, denoted $f*$, and then begins to decrease. When f is small, a small number of biased agents can dominate discourse. As f increases, opposing camps will tend to cancel each other out, although increasing f too far will introduce more polarization and decrease accuracy.

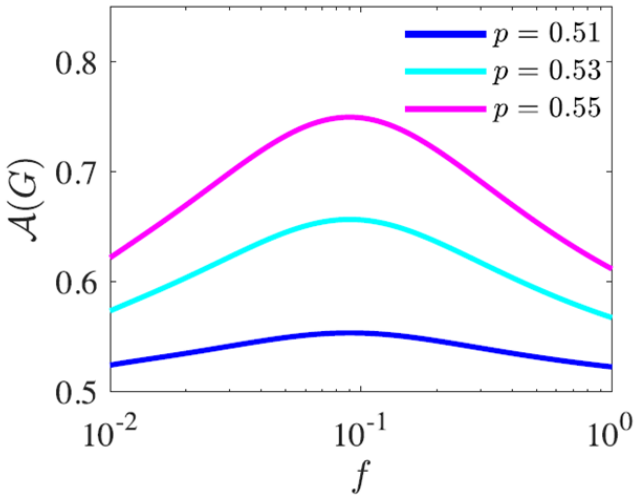

FIG. 3. Graph accuracy as a function of the fraction of biased agents in the graph.

Their results indicated, perhaps not so surprisingly, that accuracy increases with p, although for graphs with high f, accuracy increases more slowly. Additionally, a more highly-connected graph attains a higher $f*$, and a higher overall accuracy as shown in Figure 4. More connections, and more discourse, mean that the graph can discern the truth more easily while also filtering out biased news.

Either Jonathan Heidt or Jordan Peterson (or maybe both) said that truth does not lie at one end of the political spectrum, and it is at our own peril to assume so. No one has the complete picture, but when we have discourse from two sides, we get closer to it. Heidt's book, *The Coddling of the American Mind*, began with an amusing (and completely fictional) account of travelling to the Greek sage Misoponos, who presented them with three great "untruths," the third of which was "Life is a battle between good and evil people." At the risk of being overly dramatic and philosophical, I'm going to claim that this paper reminds us that we should engage with people who don't agree with us, and that there can be constructive space for disagreements even within friendships

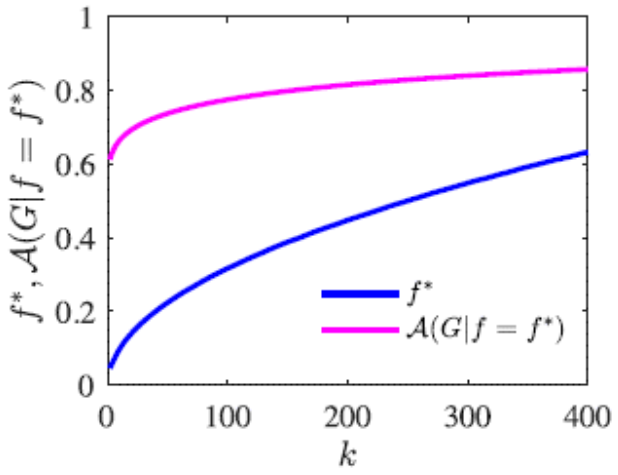

FIG. 4. Optimal f and corresponding accuracy as a function of the node degree k

III. EXPERIMENT SUMMARY

The github repository containing the code I wrote is located at this repository

Similar to Sikder et. al, I modelled a graph of n agents with edges representing friendships on the social media website. Time evolves in discrete time steps, and every agent is making an estimate of the value of X based on the information that they have been given. Information comes in packets known as Posts, which have two attributes: their bias and interest value. The interest value represents how intrinsically captivating a post is (in the real world, how well it's written, filmed, recorded, or otherwise presented) while bias represents what the post indicates about X; users will base their estimate of X based on the biases of the posts that they have consumed. Both bias and interest value are real numbers in the range $[0, 1]$

Unlike in Sikder, their estimate can take on any value in the range $[0, 1]$, and agents *can go offline*, a state in which they will neither consume nor generate information. Information is passed directly between neighbors in the graph, but it is also highly influenced by an outside agent that can pass information between non-neighbor nodes depending on some algorithm. This outside agent will be referred to as the Company, and is meant to represent a social media website that is trying to recommend content to agents that will maximize their time online.

Everyone has a *notification inbox* and a *feed*. The former holds posts which are sent directly from friends (neighbors in the graph) while the latter holds content that the Company recommends to them. The summary of the process is given below. For every time step:

1. Determine who will be online based on content viewed in the previous time step
   - Everyone will begin online for the first time step
   - Agents who are offline will have certain conditions for coming back online; the details will be discussed later
2. Determine which agents will make a post in this time step
   (a) Agents will make a post with fixed probability $\mathcal{A}$, unique for every agent
   (b) Agents make posts which have leanings equal to their own, plus a (gaussian) noise term
   (c) The interest value of a post is randomly and uniformly distributed in $[0, 1]$ for all created posts regardless of the agent that made them
   (d) Agents' posts are sent directly to their friends notification inbox, but *not* into the feed of any other agent
3. Have the Company will determine what shows up in the agents' feed
4. Have every online agents read all posts in their notification inbox, and the first k items in their feed
   (a) k is determined randomly at every time step for every agent, although every agent has an underlying parameter that determines the average value of k
   (b) Adjust their estimates of X (leaning) using a weighted average of the content they have consumed up to this point
5. Repeat

This series of steps was meant to account for complexities not accounted for in Sikder's paper, especially the mechanisms by which agents decide whether or not to go online. I hoped to capture the process in which social media sites vie for their user's attention, accommodating their preferences to keep them interested and sometimes created echo-chambers of thought that are very difficult to break out of.

Every day experience tells me that social media intake is at least somewhat stochastic, so I defined a function to quantify the probability that agents would stay online, based on how "interesting" their content was. Define the following function:

$$\begin{aligned} g &: \mathbb{R}_0^+ \longrightarrow [0, 1] \\ g(x) &= \frac{2}{\pi} \cdot tan^{-1}\left(x - \tau + \tan\frac{\pi}{4}\right) \\ S(x) &= max(0.05,\ \ g(x)) \end{aligned} \tag{3}$$

In this case, x is a positive real number that quantifies "interest" generated by the posts a user has just seen (more on this later), and the value $S(x)$ is the probability that the user will stay online after getting x interest after perusing the social media website. τ is a predefined

constant, unique for every agent, which represents the interest required to keep the user online 50% of the time, as $S(\tau) = \frac{1}{2}$. The function $g(x)$ is only introduced to make $S(x)$ more readable.

We notice that no matter how boring the content is, there is always a 5% chance that the agent will remain online. Once an agent is offline, they will check their notifications with probability.

$$f : [0,1] \longrightarrow [0,1] \tag{4}$$
$$f(x) = min\left(1, \frac{3 \cdot \beta(x,a,b) + 3}{20}\right)$$

where x is the leaning of the stimulus, and β is a beta distribution with parameters a and b. These parameters are fully determined by the agent's leaning μ and the standard deviation s:

$$a = \frac{\mu^2(1-\mu)}{s^2} \tag{5}$$
$$b = \frac{\mu(1-\mu)^2}{s^2}$$

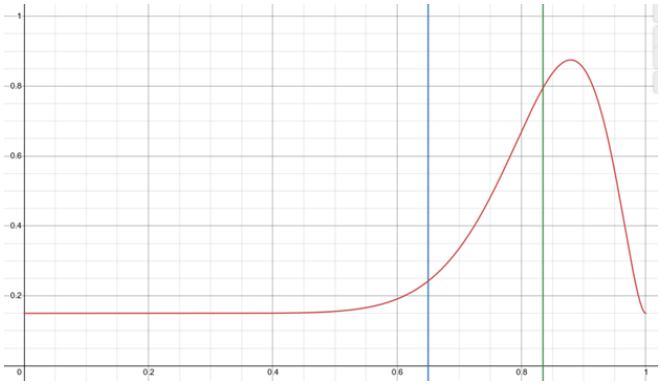

FIG. 5. Plots of Equation 4 with agent leaning $\mu = 0.65$ shown as a blue line and $\mu' = 0.835$ shown as a green line

A function of the form of a beta distribution was used in Equation 4 because I wanted to capture the idea that we engage more with news which shares our own views, or that may even have a stronger leaning but in the same direction as we do. The beta distribution tends to have probability density peak that is skewed to one side of the mean μ, as shown in Figure 5. However, this effect is not very pronounced for values of μ which are close to 0.5. To try and induce more polarization, I skewed μ using the following equation

$$\mu' = \sqrt[3]{\frac{1}{2}^2 (\mu - 0.5)} + 0.5 \tag{6}$$

We use Equation 6 before applying Equation 5 to find a and b, thus ensuring that moderate values of μ are pushed outwards. Figure 6 shows the probability distribution of μ', and we clearly see that most of the probability density falls towards the endpoints.

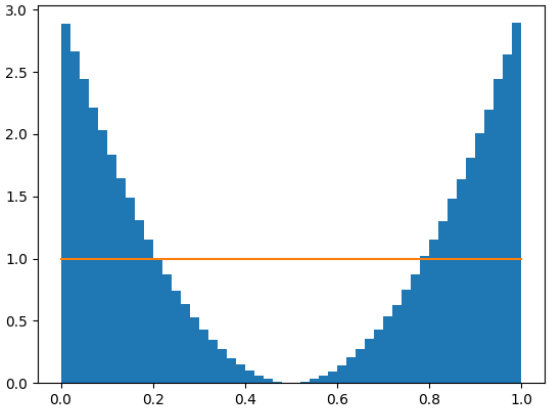

FIG. 6. The histogram of 1 million values sampled from Equation 6 (using a uniform distribution of input values, shown as an orange line). Normalized to show the approximate form of the probability distribution function

Another pleasing feature of using the beta distribution is that it can model someone "falling off the deep end." For the values of a and b that are employed in this model, the distribution strongly resembles a Gaussian for most values of μ. However, when μ approaches 0 (1), the beta function will lose all resemblance to a normal curve and instead approach ∞ for $x \to 0$ ($x \to 1$). A standard deviation of 0.9 was used, as asymptotic behavior was induced for $\mu < 0.1$ or $\mu > 0.9$ with this value of the standard deviation.

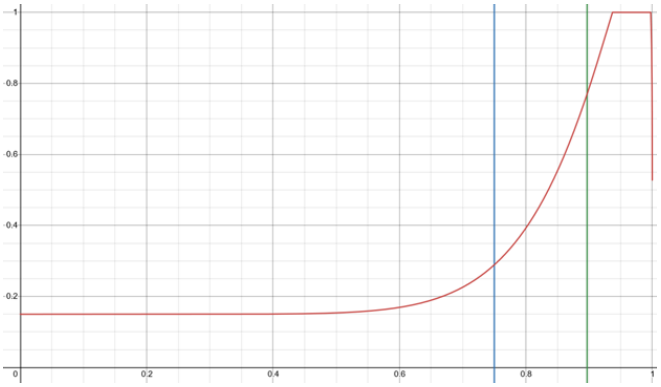

FIG. 7. Plots of Equation 4 with agent leaning $\mu = 0.75$ shown as a blue line and $\mu' = 0.897$ shown as a green line

The definition of Equation 4 caps this asymptotic behavior, but there is a nice intuition that it can capture: when your own leaning becomes markedly polarized to one side or another, you will begin to ignore most other news and instead fall into a trap of listening only to those sources which agree with you.

When reading through their feed, each agent would calculate their engagement with the post by multiplying the post's interest factor by Equation 4. This engagement was used in conjunction with the post's leaning to update the agent's beliefs. In lieu of setting a ground truth which would have enabled me to define parameters necessary for a Bayesian (or even a Naïve Bayesian) analysis, I had each agent update their beliefs by taking the engagement-weighted average of every post they had seen in the last 20 time steps, so a person's beliefs were determined by

$$\tilde{X} = \frac{\sum_{t=1}^{20} \sum_j \mathcal{E}_{tj} \times \mathcal{L}_{tj}}{\sum_{t=1}^{20} \sum_j \mathcal{E}_{tj}} \tag{7}$$

Where $\mathcal{E}_{tj}$ is the engagement the user had with the j^{th} news article in the t^{th} time step and $\mathcal{L}_{tj}$ is the leaning of the post. Anything older than the 20^{th} time step was "forgotten."

IV. RESULTS

The model displays some interesting behavior, but it unfortunately fails to polarize. Figures 8 and 9 show two runs of the algorithm that agents initialized with extreme leanings but in both cases the graph returns to moderate leanings.

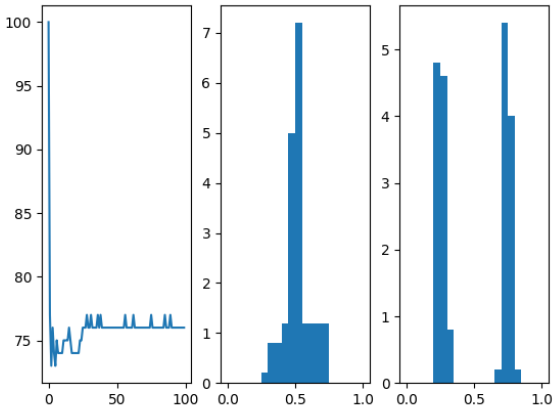

FIG. 8. Graph with 100 agents initialized to have two highly polarized groups. Trace plot of the number of people online shown at left, distribution of leanings upon completion of 100 cycles shown center, and distribution of leanings upon initialization shown at right.

It's interesting to note that in the case of depolarizing a graph with two groups who disagree with each other initially, the total number of agents online is much lower than when the graph begins with agents who mostly agree with each other. This demonstrates that the implementation of polarization did work, to a degree. When the

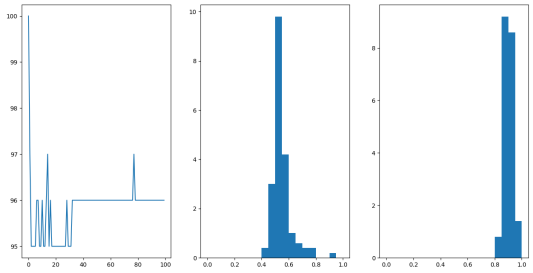

FIG. 9. Graph with 100 agents initialized to all have high estimates of X. Same layout as in Figure 8

two groups made content, the leaning of each post varied quite wildly and the Company was unable to provide as much content that matched each agent's leaning, causing them not to engage as much and go offline more frequently.

The number of people who stay online also remains and slowly climbs upwards until it reaches a steady state. This is due to an initial dearth of posts. The Company does not have enough content to provide for the agents in the initial timesteps of the algorithm. It is designed to save 3 time step's worth of agent-generated posts, which it can draw from when selecting content that it predicts will be interesting to agents. After the system stabilizes, the Company can draw from a much larger body of content to interest the agents.

I'm unsure exactly why, even after all the attempts to induce polarization, such behavior was not observed. There may factors that made polarization untenable. The effects of graph topology were quite evident in Sikder's paper, yet in this model they were randomly generated for every run.

The belief update function may have caused people to naturally tend to return to moderate beliefs. Even in the absolute limit of a gullible agent, who believes only what the last post the read states, a return towards moderate views is still observed. It would be prudent in the future to attempt to model the belief update function under a variety of stimuli, and observe how the agent responds.

One last hypothesis for the failure is that the posts generated by agents were somehow limiting the Company's options and causing it to only have moderate posts to display to its users. I find this unlike since the noise term was symmetric, but perhaps the noise term should have been dropped completely.

The hope in this experimental setup was *eliminate* polarization, via a regularization term placed on the Company, which would penalize it for high levels of bias in the graph. The form of this "tax" could have been linear (a constant times the standard deviation) and similar in form to ridge regularization in linear regression, or quadratic (a constant times the variance) which would have been analogous to LASSO regression. Via this tax, the company would have had to adopt a less aggres-

sive posting schedule, and displayed news that sometimes conflicted with an agent's own views to reduce the disparity in opinions in the graph.

V. CONCLUSION

Although the algorithm displayed some interesting behavior, it was overall unsuccessful in modelling polarization. If I were to repeat this experiment, I would first attempt to recreate Sikder's results, and build more complexity into that model. Sikder's results are fascinating, both in their agreement between analytic results and computer modelling, but also because it provides quantitative results that have a sociological intuition.

The underlying problem is that there would have been no way to account for all of these factors, even given much more time and computational power; it had a fatal flaw. The model that was presented in Section III was overly complicated, and introduced far too many complicating factors whose impact is indeterminate, and whose computational impact was large. Less emphasis should have been placed on using floating point values, and integers should have been employed. Agents should have been more similar than dissimilar, and the scope of tunable parameters for each agent should have been reduced. "A designer knows he has achieved perfection not when there is nothing left to add, but when there is nothing left to take away" - Antoine de Saint-Exupéry. If this project is repeated, it should be known that emphasis should have been on minimalism, rather than on scope.

---

[1] Taken from Illustration Chronicles.
[2] This was taken from Tim Wu's *fascinating* book, *The Attention Merchants* end of chapter 3.
[3] The article is available here.