Network Diffusion with Conflicting Parties

Abishrant Panday
abishrantpanday@college.harvard.edu

William Zhang
william_zhang@college.harvard.edu

June 6, 2021

Abstract

Expanding on the core literature of single-idea propagation, Sun et al. [1] and Zhang et al. [2] explore competing information diffusion in social networks. In this work, we extend [1] to create a model for network propagation with two parties and explore its dynamics under different conditions. We find that in the Facebook ego dataset, it is possible to determine the relative value of the node of highest degree, $v_{\max}$, with respect to a set of random nodes in the graph. Moreover, we show the impact of distance to $v_{\max}$ on information spread and illustrate a strategy to suppress information. Finally, we consider two strategies within the budget-constrained model and show the effectiveness of seeding a large number of nodes with low degree distributed over the graph instead of picking fewer nodes with high degree.

1 Introduction

Social networks, which are composed of the interpersonal relationships within a group of individuals, play a crucial role in the spread of information. Often, people are influenced by others in their adoption of a novel idea—whether it's a new smartphone or a medical advancement—and the study of the network processes which are critical in these interactions has been a growing part of the social sciences literature [3].

Specifically, an important avenue of research in the field concerns ways of maximizing this spread of information within a social network. Empirically, these studies have focused on issues ranging from optimal viral marketing plans for brand awareness [4] to the diffusion of agricultural technologies in developing countries [5].

Much of the research in this area revolves around information diffusion through optimally informing agents, where the focus is on transmission of a single idea. However, many real-world instances of marketing and information-spread tend to involve multiple competing parties—such as clothing manufacturers trying to sell a shoe or Samsung and Apple vying for market share. In this paper, we thus explore the network dynamics and optimal seeding strategies in the presence of two competing agents. We focus on a specific information propagation model, discussed in Section 3, and utilize properties of the underlying graph to inform our heuristics. We then investigate the impact of constraining these strategies, such as through budget limitations, and provide insight into methods of *suppressing* information as well.

The paper is organized as follows: Section 2 presents a review of current literature relevant to the work. This is followed by a description of our model in Section 3 and important results in Section 4. We conclude and present future directions for our model in Section 5.

2 Related Literature

Work by Kempe et al. in 2003 [6] showed that it is computationally unfeasible to identify individuals who are best able to maximize information diffusion in a network. Their work, in addition to noting that the problem of selecting the most influential nodes in a network is an NP-hard, also explored the performance guarantees of approximation and natural greedy heuristics for this problem [7]. Due to the aforementioned intractability, many papers are focused on utilizing network characteristics that are related to influence, such as eigenvector centrality and degree centrality [8]. Although these characteristics are useful for the information diffusion problem, determining the values within real networks is costly, as shown by Breza et al. [9], who developed methods to determine network structures without needing full information.

As mentioned in the introduction, work by Domingos and Richardson [4] investigated how to best utilize *viral marketing* within a network. Here, advertisers prioritize individuals who have the most influence in order to potentially trigger a cascade that leads many people to try the product without the high costs of *direct marketing* to each person. The paper works in the paradigm where the joint distribution over all nodes' actions is specified and presents presents heuristics for identifying nodes with a large influence on the network.

Work by Morris in 2000 [10] focused on analyzing general local interaction systems, where there are an infinite number of players who interact with their neighbours and can choose a binary option to play. In this framework, Morris characterized the conditions under which contagion—when a behaviour spreads to the entire population—occurs.

In our paper, we choose a modification of a specific propagation method: the SIR model [11], which is an epidemic model that subdivides a population into different compartments based on their "infection" by an idea and simulates dynamics as members move through the various compartments. This is similar to the framework utilized by Akbarpour et al. [12], who explored the impact of seeding in information diffusion and showed that the strategy of randomly seeding $s + x$ individuals can lead to greater diffusion than choosing the optimal s.

A paper that works in a similar paradigm to our model is that by Sun et al. [1], which explored competing information diffusion in social networks and the resulting dynamics. Their research focused on the influence of different information spreading probabilities, changing densities of 'seed' nodes, and different network structures.

Finally, a paper by Zhang et al. from 2013 [2], also considered a competitive maximization problem with the goal of choosing seed nodes to increase information spread. In their model, nodes had preferences that were explicitly determined by a probability distribution, which then determined the information to be adopted. The main result was a heuristic algorithm based on a random walk on the network, where the most influential nodes were identified by tracing information back based on the random walk to find probable origins.

3 MODEL

3.1 Information Propagation

We will follow the framework presented in Sun et al. [1], which resembles a modified SIR model.

Specifically, the model consists of a graph $G = (V, E)$ and two sets of information, A and B. The graph has some initial "seed" nodes, with states S_A, S_B, which diffuse information A, B respectively to their neighbours. The remaining nodes are in an ignorant state I, where they do not know any information. The final set of nodes are "recovered," meaning they can no longer spread information, but are not ignorant. These are R_A, nodes which know A but not B, R_B, which know B but not A, and R_{AB}, which know A and B. The end of the diffusion process is when there are no spreading nodes in states S_A, S_B, meaning all nodes will be one of $\{I, R_A, R_B, R_{AB}\}$.

The diffusion process, illustrated graphically in figure 1, occurs as follows:

- At every time step, each person in state S_A either informs their neighbours about A with probability p_A or decides to stop transmitting information (and thus enters state R_A) with probability q_A. The analogous quantities for information B are p_B, q_B. We note that the transition between spreading and recovery states reflects the notion that people tend to not share news they believe to be "old."
- If a node in state I is exposed to both A, B simultaneously, then it randomly accepts either state with probability $\frac{1}{2}$.
- If a node is in the spreading state of S_A, S_B then they are unaffected by others' information.
- Finally, if a node in recovered state R_A is exposed to information B then they transition to state R_{AB} with probability r_A. The analogous probability for the $R_B \rightarrow R_{AB}$ transition is r_B.

The model omits direct transitions $I \rightarrow R_A, R_B$ and $S_A, S_B \rightarrow R_{AB}$ from the Sun et al. [1] paradigm.

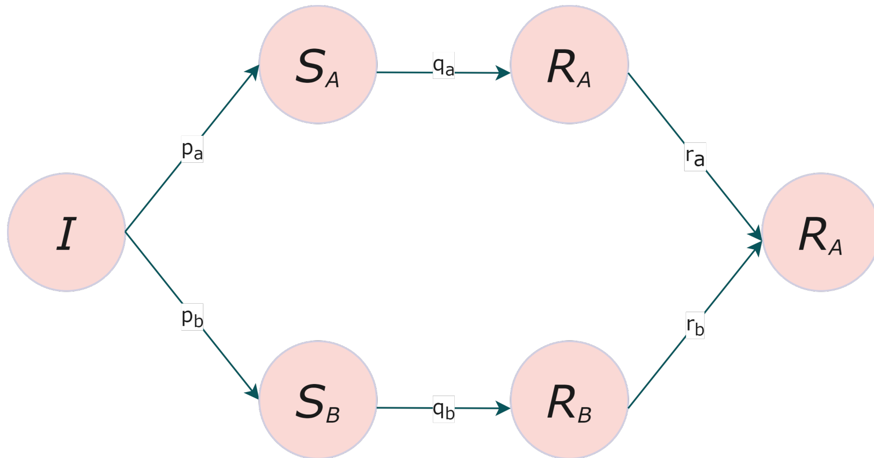


Figure 1: Possible state transitions and their probabilities given the conditions in Section 3.1

3.2 Budget Constraints

In addition, we consider the situation where the competing agents, whose goal is to maximize the spread of either A or B in the network, work with a constrained amount of seed nodes. Specifically, we implement a cost to creating a seed node and have agents operate with a limited budget. We assume that the cost to make $v \in V$ a seed node is equal to its degree, $\deg(v)$. This reflects the real-world situation where agents are looking to "sponsor" nodes to propagate information and more popular nodes are more costly to use as advertisers.

3.3 Experimental Setup

Unless stated otherwise, experiments are run on the Facebook-ego graph dataset [13] [14], which is hand-labeled data that consists of Facebook friends lists. This is an undirected graph, due to the nature of Facebook friendship, with $4,039$ nodes, corresponding to users, and $88,234$ edges, a visual representation of which is Figure 2. The node degrees in the Facebook-ego graph follow a power law, instead of being regular. In particular, there is a node $v_{\max}$ with degree $1,045$.

The transition probabilities are fixed as well, unless indicated otherwise:

$$\begin{bmatrix} p_A, p_B \\ q_A, q_B \\ r_A, r_B \end{bmatrix} = \begin{bmatrix} 0.2, 0.2 \\ 0.1, 0.1 \\ 0.05, 0.05 \end{bmatrix}$$

Let the number of nodes be $N = 4039$. At $t = 0$, some number of seed nodes $S_A(0), S_B(0)$ are initialized based on each experiment condition. There are no nodes in the resting state: $R_A(0) = R_B(0) = R_{AB}(0)$. All remaining nodes are in the ignorant state: $I(0) = N - S_A(0) - S_B(0)$.

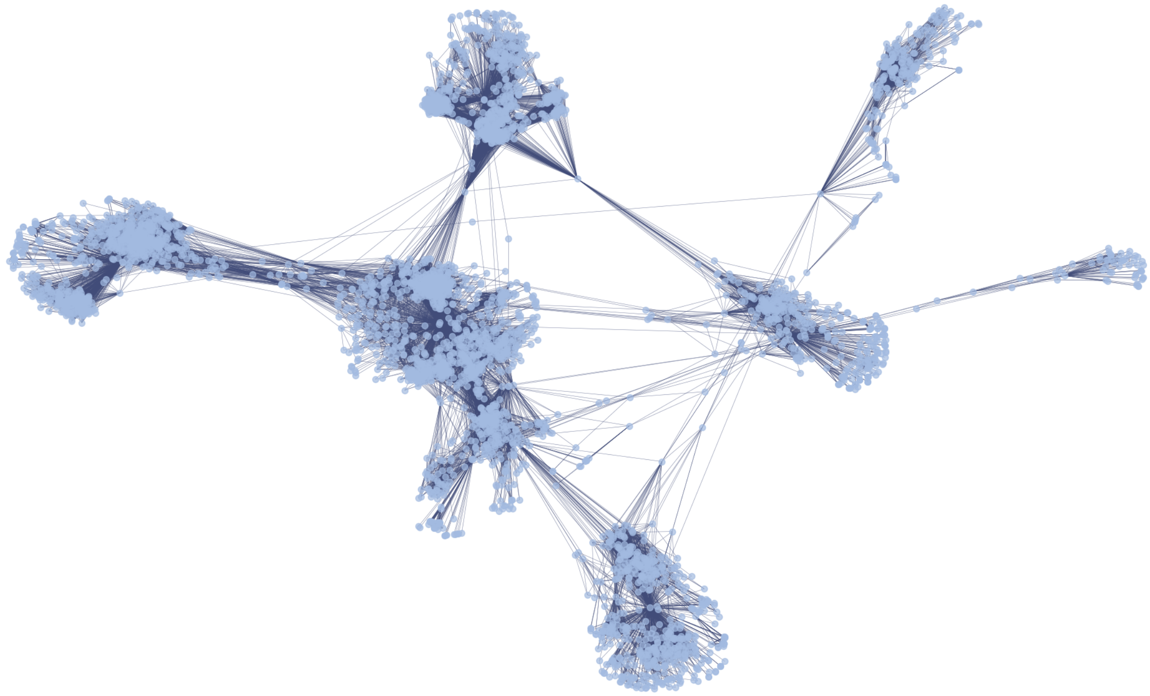


Figure 2: Visual representation of the Facebook-ego graph. Edge lengths are arbitrary.

4 RESULTS

Each simulation consists of 20 runs, from which the data is then aggregated. Each run spans 40 timesteps, which is enough to ensure convergence for all scenarios. At the end of each simulation, the overall winner between A, B is determined by whichever of $\frac{R_A}{N}$ or $\frac{R_B}{N}$ is larger.

4.1 $\mathbf{V}_{\max}$ vs Random Seeding

A was given a single seed at $v_{\max}$ and B had k randomly chosen seeds for $k \in [5, 20]$. We found that for $k < 9$, A was the winner while for $k > 9$, B was the winner. The results are shown in Figure 3 for $k \in [6, 12]$, where the data indicates a stark difference in the $\frac{R_A}{N}$ and $\frac{R_B}{N}$ values between the $k = [6, 8]$ and $k = [10, 12]$ groups. Figure 3 (d-f) also illustrates that the graph had reached a steady state, with the fraction of nodes in the I, S_A, S_B states at less than 1%.

The $\frac{R_A}{N}$ and $\frac{R_B}{N}$ values converged to within 0.001, meaning the final R_A, R_B amounts differed by at most 4 nodes, for $k = 9$ and this case was further analyzed for 100 runs. This is illustrated in Figure 4, where we see that having 9 random seeds is approximately equivalent to having one seed at $v_{\max}$. These results show that it is possible, in the Facebook ego dataset, to quantify the value of a node's information propagation abilities with respect to any set of random nodes in the graph.

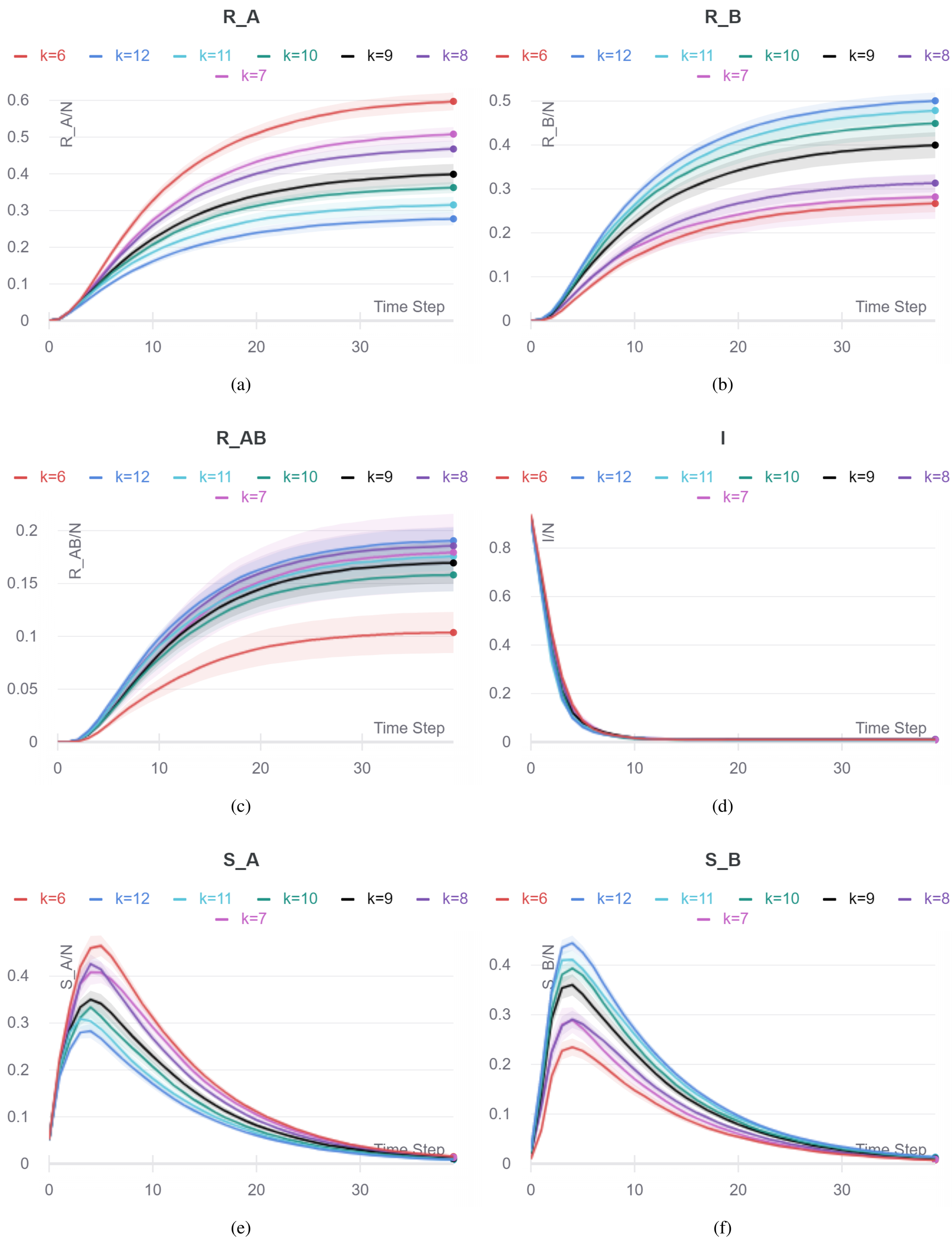
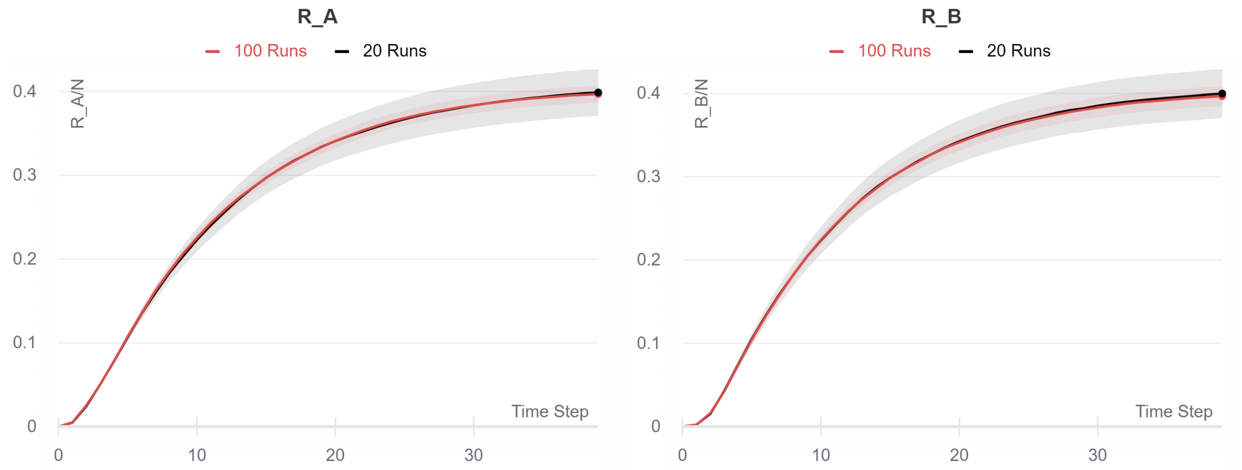


Figure 3: Fraction of nodes in each state at a given timestep for $v_{\max}$ vs random seeding, with standard error.



(a) (b)

Figure 4: Fraction of nodes in R_A and R_B for $v_{\max}$ vs 9-random seeds and $20, 100$ runs, with standard error.

4.2 Single Seed Start

Both A and B were given a single, randomly chosen seed node in the graph. In addition to the above diffusion metrics, each seed node's distance from $v_{\max}$, denoted D_A, D_B, was also computed. There were two key findings:

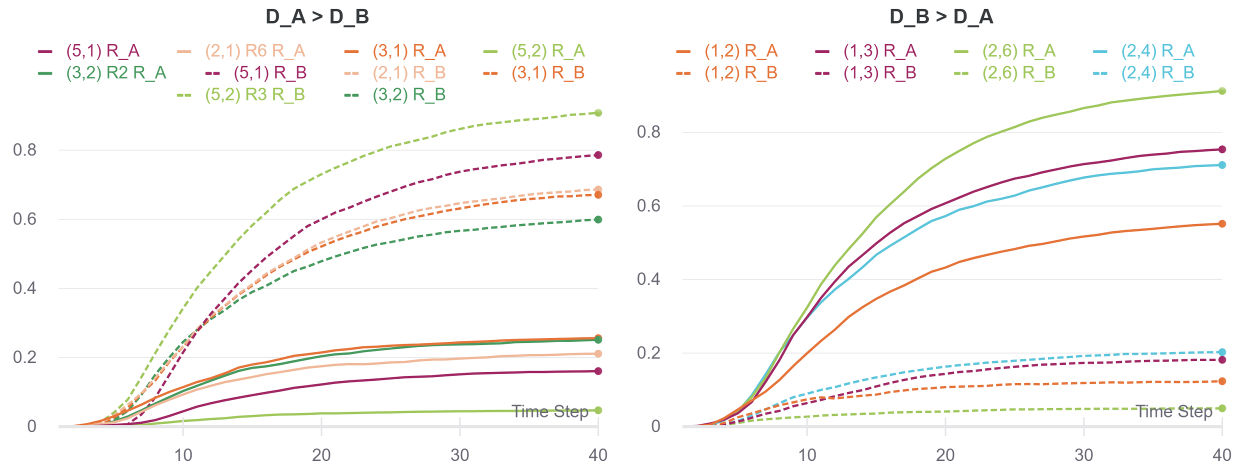
- If $D_A = D_B$, meaning the two seed nodes were at the same distance from $v_{\max}$, then the winning party was determined entirely by whichever node had higher degree.

- If $D_A \neq D_B$, then the node with the shorter path to $v_{\max}$ won consistently, even if it had smaller degree.

The second result is illustrated in Figure 5, where the runs were separated into those where seed A was closer to $v_{\max}$ and where B was closer. In each case, there was a divide in the final R_A, R_B curves where $R_A > R_B$ if $D_A < D_B$ and vice versa. Moreover, across experiments where the starting distances (D_A, D_B) were the same, it was found that the magnitude of the final R_A, R_B values increased with increasing degrees of the seed nodes, while still following the trend of the second result above.

4.3 $\mathbf{V}_{\max}$ vs Strategic Seeding

A was given a single seed at $v_{\max}$ and B was allocated three seeds at the second, third, and fourth highest-degree neighbors of $v_{\max}$. Denote the neighbours as v_1, v_2, v_3 respectively. In the Facebook ego graph, these nodes had degrees $\deg(v_1) = 349, \deg(v_2) = 347, \deg(v_3) = 254$. The results are shown in Figure 6, where, over all runs, the average end-state of information propagation had $R_B > R_A$.



(a) (b)

Figure 5: R_A, R_B curves for randomized single-seed experiment. Separated into (a) $D_A > D_B$ and (b) $D_B > D_A$.

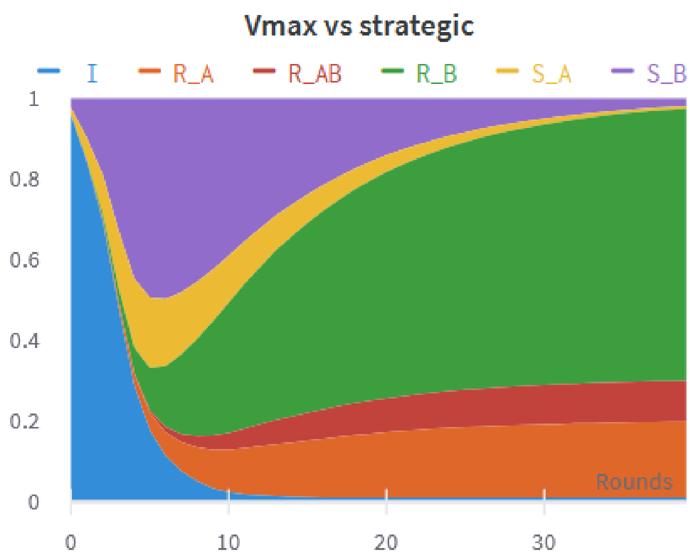


Figure 6: Percent area plot of node states in the $v_{\max}$ vs strategic seeding experiment.

4.4 Budget Constraints

Each node $v \in V$ was given cost equal to its degree $\deg(v)$ and each experiment consisted of multiple runs at a given budget $\mathcal{B}$. Agents followed one of two strategies in selecting seed nodes:

- **High:** Agents greedily selected $n_H \in [3, 6]$ nodes v_i such that for each node $\deg(v_i) \sim \frac{\mathcal{B}}{n_H}$.
- **Low:** Agents greedily selected $n_L \in [20, 50]$ nodes v_i such that for each node $\deg(v_i) \sim \frac{\mathcal{B}}{n_L}$.

To determine reasonable budgets, the degree distribution of the graph, shown in Figure 7, was calculated. The sum of all degrees was $176,468$ and the average degree was 43.69.

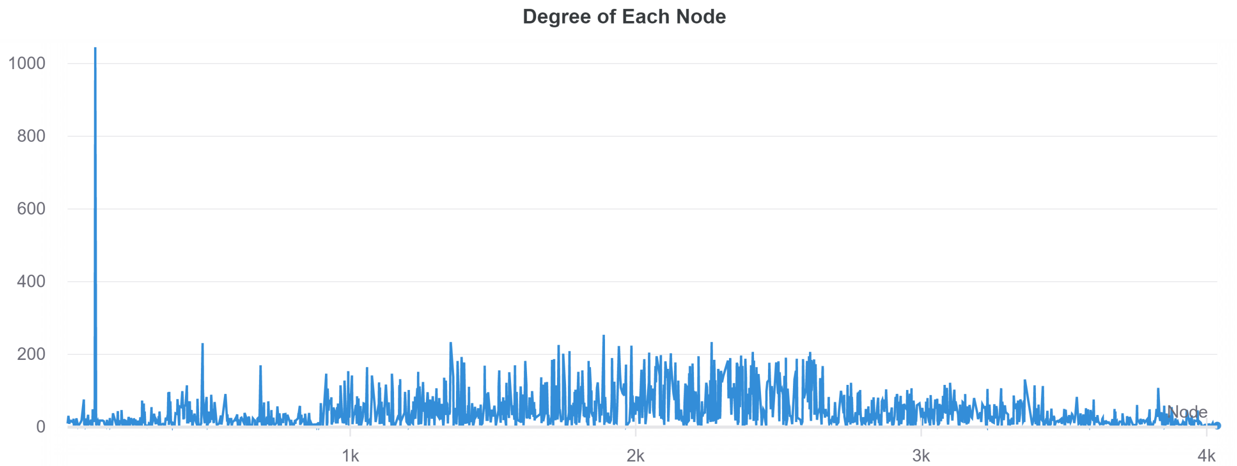


Figure 7: Distribution of node degrees; x-axis runs through all nodes.

Experiments were run for $\mathcal{B} = 200k$, $k \in [1, 10]$. A was seeded according to **High** and B was seeded according to **Low**. It was found that for each budget $\mathcal{B}$, the number of nodes in state R_B was larger than the number of nodes in state R_A, meaning adopting **Low** was an advantageous strategy. Figure 8 illustrates the $\frac{R_A}{N}$ and $\frac{R_B}{N}$ curves for several combinations $(\mathcal{B}, n_H, n_L)$.

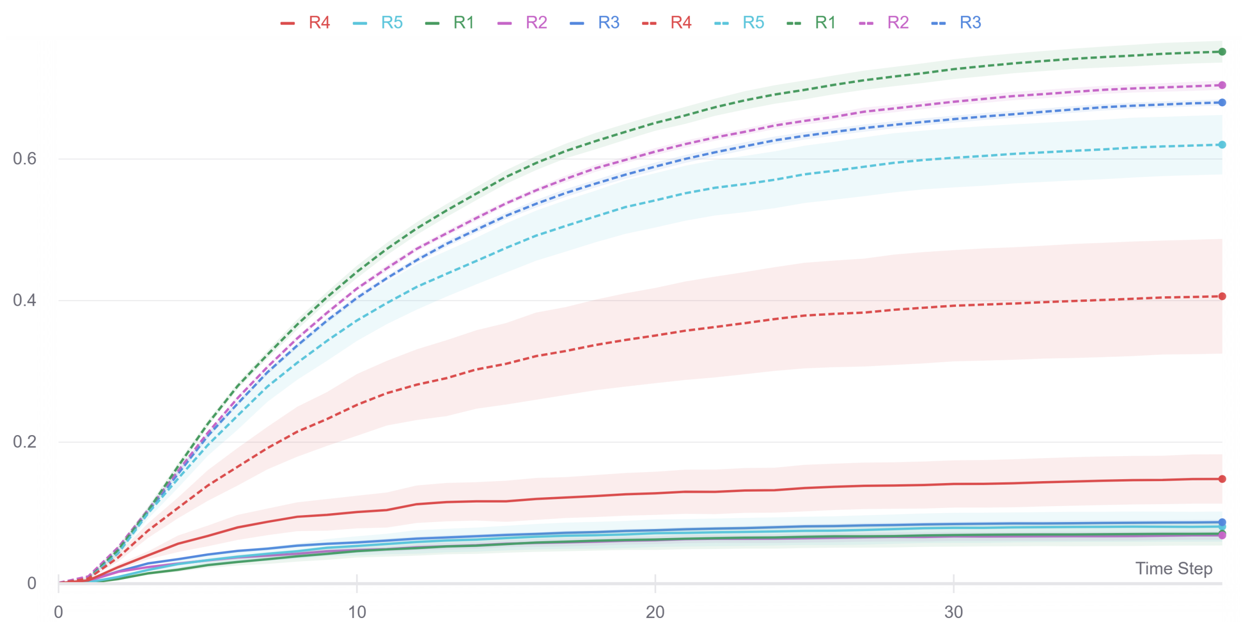


Figure 8: Runs are labeled $R1 = (2000, 3, 50)$, $R2 = (2000, 6, 50)$, $R3 = (1400, 5, 40)$, $R4 = (1000, 4, 20)$, $R5 = (600, 6, 30)$, with solid lines for R_A and dashed lines for R_B.

5 Discussion and Future Work

In this paper, we proposed a model of network diffusion in the presence of two competing parties and looked at the effect of different seeding strategies on information propagation in the Facebook ego graph. We showed in Section 4.1 that in the dataset, in expectation the propagation effect of $v_{\max}$ was equivalent to that of 9 random nodes. This method provides a method by which the relative value of a node can be determined with respect to the average node in the graph. Future experimental work in this area should determine whether, for different graph structures and node degree distributions, an exact equivalence still holds. Theoretically, working within the framework of Sun et al. [1] and Zhang et al. [2], future work would entail considering more regular graph structures in order to find provable bounds on the existence of such a knife-edge property.

In Section 4.2, we showed that the impact of a single seed depended most on its distance from $v_{\max}$. However, is is unclear how large of a role the degree distribution, as seen in Figure 7, plays in this result. Specifically, an extension should look at graphs without extreme outliers in order to see whether this effect still exists and its relative importance compared to seed node degree.

Finally, Sections 4.3 and 4.4 showed results that could be useful in real-world marketing applications. Specifically, 4.3 illustrated the efficacy of choosing nodes around the opposition's seed(s), even when the total degree of the surrounding nodes is less than the opposition degree. 4.4 then indicates that at a given budget, it is better to pick many nodes distributed over the graph with low degree instead of picking fewer nodes with high degree. This section provides the most avenues for future work. The **High** and **Low** strategies were relatively unsophisticated and could be improved to provide better diffusion results. It would also be informative to train a model that choose the best nodes for a given budget, although this is unfeasible in the real-world, in order to get a baseline for relative effectiveness. Moreover, defining a node's cost as its degree is a strong assumption and different heuristics should be explored, including other measures of centrality. Finally, the results ignore the outside factors that may make **Low** a worse strategy, including the cost of actually finding and partnering with many smaller agents as opposed to more established, central agents and the "trust" placed in smaller agents. The second point in particular could be addressed by augmenting the model to give each node a propagation factor proportional to its degree.

References

[1] Q. Sun, Y. Li, H. Hu, and S. Cheng, "A model for competing information diffusion in social networks," *IEEE Access*, vol. 7, pp. 67916–67922, 2019.

[2] B. Zhang, Z. Qian, X. Wang, and S. Lu, "Tracing influential nodes in a social network with competing information," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 37–48, Springer, 2013.

[3] E. Breza, A. Chandrasekhar, B. Golub, and A. Parvathaneni, "Networks in economic development," *Oxford Review of Economic Policy*, vol. 35, no. 4, pp. 678–721, 2019.

[4] M. Richardson and P. Domingos, "Networks in economic development," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 61–70, 2002.

[5] L. Beaman, A. BenYishay, J. Magruder, and A. Mobarak, "Can network theory-based targeting increase technology adoption?," *National Bureau of Economic Research*, no. w24912, 2018.

[6] D. Kempe, J. Kleinberg, and Tardos, "Maximizing the spread of influence through a social network," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, 2003.

[7] D. Kempe, J. Kleinberg, and Tardos, "Maximizing the spread of influence through a social network," *Theory OF Computing*, vol. 11, no. 4, pp. 105–147, 2015.

[8] B. Golub and E. Sadler, "Learning in social networks," *Available at SSRN 2919146*, 2017.

[9] E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan, "Using aggregated relational data to feasibly identify network structure without network data," *American Economic Review*, vol. 110, no. 8, pp. 2454–2484, 2020.

[10] S. Morris, "Contagion.," *The Review of Economic Studies*, vol. 67, no. 1, pp. 57–78, 2000.

[11] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics.," *Proceedings of the royal society of London. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.

[12] M. Akbarpour, S. Malladi, and A. Saberi, "Just a few seeds more: value of network information for diffusion.," *Available at SSRN 3062830*, 2020.

[13] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks.," in *NIPS*, vol. 2012, pp. 548–56, Citeseer, 2012.

[14] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection." http://snap.stanford.edu/data, June 2014.