Markov Chain Model Representation of Information Diffusion in Social Networks

Kagan Alexander

University of Michigan

April 2023

1 Introduction

Information Diffusion is the process by which information spreads through a network. Social networks are naturally modeled as graphs of agents, in which the agents are represented by vertices that are connected by edges if the agents can share/communicate information. Diffusion processes have been extensively studied from the non-Markovian perspective in the network analysis literature, e.g. see Kempe et al. [2003], Jackson et al. [2017]. Contrary to all previous literature, this work aims to model the information diffusion process as a Markov Chain where at each time point a randomly chosen "infected" person independently attempts to infect all its uninfected neighbors. Despite its similarity to the famous SIR model (Radcliffe [1977]), the approach of Dennis et al. [2022] studied in this report takes into account the topology of the social network and thus can significantly improve the spread modeling in cases when the graph structure is observed. Moreover, the described model allows considering "idea" and "anti-idea" spreaders simultaneously. In a social network context, idea-infected nodes can play the role of fake news spreaders while anti-idea-infected nodes can be thought of as users spreading the warnings that this news is fake.

This report aims to reproduce the Markov Chain infection model described in Dennis et al. [2022]. We extend the experiments described in this work with extensive simulations studying the behavior of the model in the case of complete graphs. By using complete graphs we eliminate the problem of the unfair initial allocation of "idea" and "anti-idea" spreaders and thus allow to run more complex and reliable experiments.

2 Proposed information diffusion model

Consider a directed graph $G = (V, E)$ where each node is either (1) infected with idea ϕ, (2) infected with an anti-idea $\neg\phi$, or (3) indifferent to this idea $\perp\phi$. Denote each of these sets at times $t = 0, 1, \ldots$ as $A_\phi^t, A_{\neg\phi}^t$, and $A_{\perp\phi}^t$ respectively, so that by definition

$$V = A_\phi^t \sqcup A_{\neg\phi}^t \sqcup A_{\perp\phi}^t \quad \forall t \in \mathbb{N}_0.$$

We assume the process starts with the known initial configuration $A_\phi^0, A_{\neg\phi}^0$, and $A_{\perp\phi}^0$ and then proceeds for $t \geq 1$ as follows

1. Choose a random node $v \in A_\phi^{t-1} \sqcup A_{\neg\phi}^t$

2. Broadcast a message in favor of ϕ if $v \in A_\phi^{t-1}$ or in favor of $\neg\phi$ if $v \in A_{\neg\phi}^{t-1}$ to **all** v's neighbors, i.e. $N(v) := \{u : (v, u) \in E\}$.

3. For each node u receiving a message in favor of ϕ (resp. $\neg\phi$), the probability to adopt it, i.e. probability that $u \in A_\phi^t$ (resp. $u \in A_{\neg\phi}^t$) is:

    (a) λ, if $u \in A_{\perp\phi}^{t-1}$, i.e. u stops being indifferent,

    (b) μ, if $u \in A_{\neg\phi}^{t-1}$ (resp. A_ϕ^{t-1}), i.e. u changes its opinion to the opposite one,

    (c) 1, if $u \in A_\phi^{t-1}$ (resp. $A_{\neg\phi}^{t-1}$), i.e. u remains in its previous state.

The process stops at the first time T when either ϕ or $\neg\phi$ infected all the nodes. In other words, the stopping time of the process is:

$$T = \inf\{t : V = A_\phi^t \text{ or } V = A_{\neg\phi}^t\}.$$

Remark. If we consider the global model with $3^{|V|}$ states (all possible configurations of nodes' labels), the resulting process will be clearly a Discrete Time Markov Process (DTMC). However, explicit computations within this model will be often intractable since the number of states grows exponentially with the number of nodes. Therefore, all experiments in the next section will actively use Monte Carlo simulations.

3 Experiments

All experiments can be found in the GitHub repository via the link:

    https://github.com/AlexanderKagan/MC-Representation-of-Information-Diffusion

Consider a complete graph G with 10 nodes with one node initially infected with the idea (blue), one infected with anti-idea (red), and other nodes being indifferent (see Fig. 1).
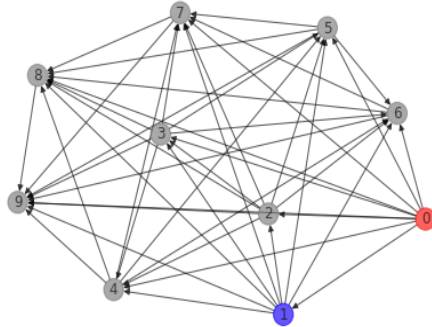

Figure 1: Complete graph with 10 nodes

3.1 Dependency of stopping time from (λ, μ) parameters

In this simulation, we run the algorithm on the graph in Fig. 1 and note the value of T, i.e. the number of messages needed to make all the nodes infected with either idea or anti-idea. The heatmap in Fig. 2 represents the log-number of messages averaged over 10^4 MC simulations for each pair of (λ, μ) in the grid $[0.1, 02, \ldots, 0.9]^2$ Using an educated guess, we decide to check if the stopping time solely depends on the ratio λ/μ. From the right plot in Fig. 2, it is clear that the ratio is indeed playing an important role in the estimation of $\mathbb{E}T$.

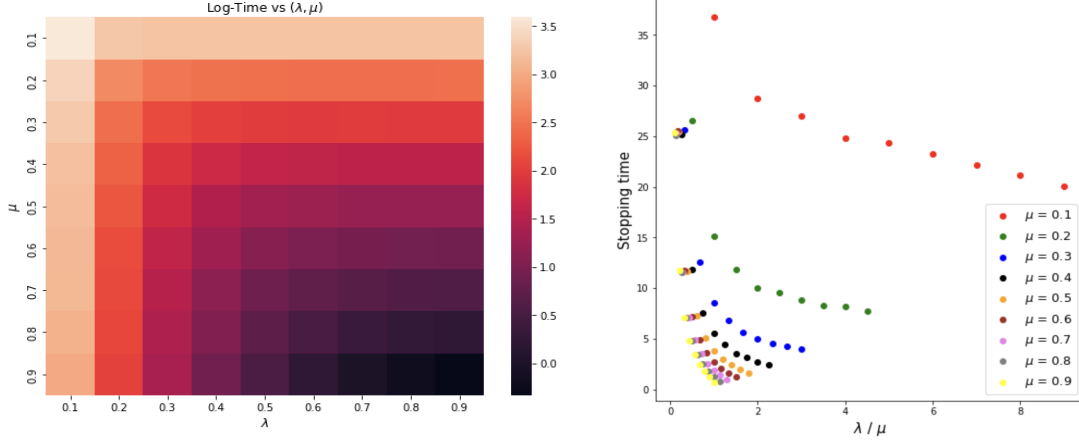Figure 2: We see that λ and μ have a roughly similar effect on the average time even though the μ parameter stops playing any role after no indifferent nodes are left.

3.2 Depnedency from the proportion of idea and anti-idea seed nodes.

Denote the initial proportions of idea spreaders and anti-idea spreaders in the seed by I and A respectively. Then, from the symmetry of a complete graph, we intuitively expect:

$$P(\text{Idea makes a full spread}) \approx \frac{I}{A + I}$$

Let us check our intuition experimentally. For each feasible pair (I, A), s.t. $I + A \leq 1$ (remaining nodes are indifferent), we run the infection model simulation 10.000 times till all nodes are infected with either idea or anti-idea and note the proportion of times when idea wins, i.e. infected all the nodes. The obtained heatmap is presented in Fig. 3. To give even more support to our hypothesis, we plot the Monte Carlo estimates of the probability that the idea makes a full spread versus the desired quantity $I/(I + A)$.

3.3 Dependencey from the size of complete graph

Let the propagation rates be $\lambda = \mu = 0.5$. For complete graphs with $|V| = 2^n$ where $n = 4, 5, \ldots 9$, run 10^4 simulations starting and one idea and one anti-idea node, i.e. $|A_\phi^0| = |A_{\neg\phi}^0| = 1$. For each graph note the average stopping time T and its standard deviation. Using an educated guess we came to the conclusion that $\mathbb{E}T(n) \sim \log_2 n$. Fig. 4 supports our hypothesis. Moreover, it is interesting to see that $\text{Var}\, T(n)$ remains approximately the same for all n. This can likely be explained by the fact the proportion of the infected nodes during each step is the same (0.5) for all n.

References

Louise A Dennis, Yu Fu, and Marija Slavkovik. Markov chain model representation of information diffusion in social networks. *Journal of Logic and Computation*, 32(6):1195–1211, 03 2022. ISSN 0955-792X. doi: 10.1093/logcom/exac018. URL https://doi.org/10.1093/logcom/exac018.

Matthew O. Jackson, Brian W. Rogers, and Yves Zenou. The economic consequences of social-network structure. *Journal of Economic Literature*, 55(1):49–95, March 2017. doi: 10.1257/jel.20150694. URL https://www.aeaweb.org/articles?id=10.1257/jel.20150694.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. page 137–146, 2003. doi: 10.1145/956750.956769. URL https://doi.org/10.1145/956750.956769.

J. Radcliffe. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(1):85–87, 1977. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346882.
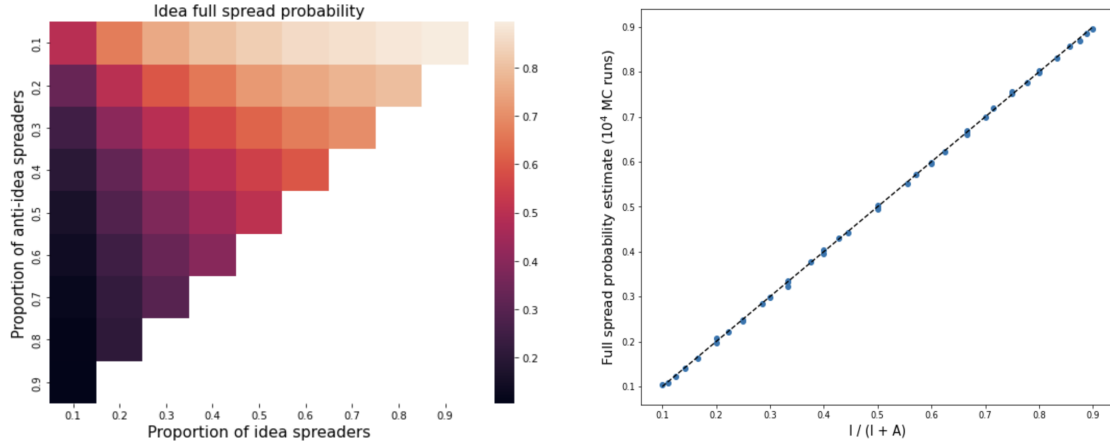
Figure 3: We see that indeed the proportions of idea wins are very similar to the proportions of idea infected seed nodes to the overall number of infected nodes
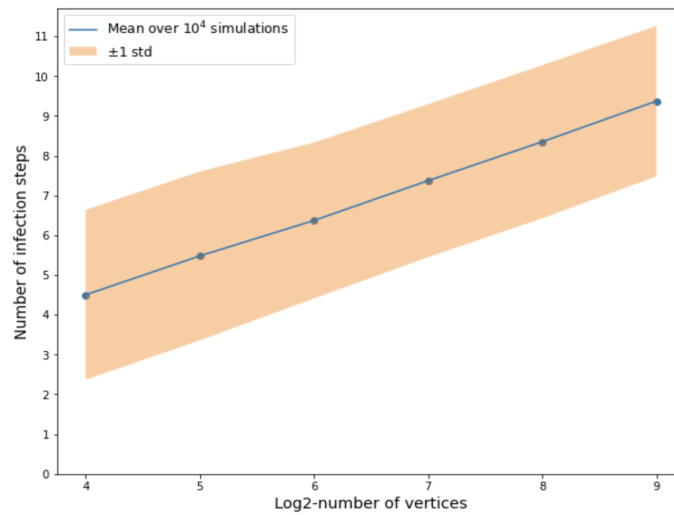

Figure 4: Dependency between $\log_2$ of graph size and average stopping time T