Analysis of Influence Propagation and Data Loss in Social Networks

BCSE202L – Data Structures and Algorithms

Abishek Devadoss 22BCE1477
C Vinston Jose 22BCE1899

Abstract – Social networks have become a very widely explored field in the recent times with the advent of various social networking applications. We aim to research about this field by looking into the concept of influence propagation and data loss. In this research, we investigate the spread/flow of infections in social networks, using the independent cascade model. The study focuses on understanding the dynamics of infection propagation and developing effective strategies for controlling the spread. We employ graph theory and probabilistic modelling to simulate the diffusion process. By evaluating various scenarios, we analyse the impact of different parameters on the infection spread and assess the efficacy of containment strategies.

Key terms: Independent Cascade Model, Social Networks, Infection Spread, Graph Theory, Probabilistic Modelling, Containment Strategies.

1. INTRODUCTION

Social networks were a very new term to the society, which is now used on a daily basis on the advent of multiple applications like Facebook, WhatsApp, Instagram, etc. They have managed to connect the entire world into a great web through the internet. These platforms have allowed the communication of individuals from different parts of the world with minimal effort from both sides.

These networks are of great importance to all fields, such as business, marketing, medical and agricultural innovations. By analysing the influencing manner among users and the spreading manner of influence based on social networking, the following advantages can be obtained: 1) in terms of sociology, it is helpful to understand people's social behaviours; 2) in terms of public services, it is helpful to provide a theoretical basis for public decision making and public opinion guidance; 3) in terms of country, it is also

helpful to promote national security, economic stability, economic progress, and so on. As a result, social influence analysis in social networks has important social significance and application value. Hence it is extremely important for research to flourish in this field of social sciences.

Our research mainly focuses on one idea: influence propagation and data loss. Influence propagation refers to the manner in which information is transferred or data is spread in a social network. Understanding the dynamics of influence spread in social networks can in turn be implemented on fields like epidemiology, sociology, and computer science.

It is of great significance to the medical field as it can prove to act as a model to study disease/infection spread among a population. Such models can simulate to a great precision with which diseases spread through populations under different circumstances, thus resulting in the formation of an epidemic. We will explore more into that in this research.

2. LITERATURE STUDY

In the realm of epidemic spreading models, extensive research has been conducted to understand the dynamics of infections in various networks. Notable studies include the work by Pastor-Satorras and Vespignani [1], where they introduced the concept of percolation theory to study the behaviour of epidemics on complex networks. Their findings laid the foundation for analysing the vulnerability of networks to infectious diseases, emphasizing the critical role of network topology. These studies were not only beneficial to the study of epidemics and infection spread, but also helped in providing valuable insights for understanding spread of information within social networks like social media platforms. The study on network topology offers parallels to the dynamics of influence spread in social networks.

Another significant contribution comes from Kempe, Kleinberg, and Tardos [2], who extensively studied the independent cascade model in the context of information diffusion. Their research focused on understanding the rapid spread of information in online social networks. They proposed algorithms to identify influential nodes for maximizing information diffusion, highlighting the relevance of early adopters in accelerating the dissemination process. From their studies it was quite evident that more influential individuals should be targeted to bring about a more widespread change or spread in any

information. By looking through the views of epidemic spread, one can identify the more influential target and take measures to prevent it from being infected to avoid more damage/spread. In this case, the target could mostly be a group of individuals, like a community, a town, a village, or a city which may be located at the centre of a larger population distribution.

Recent advancements in epidemiological studies have integrated real-world data into epidemic models. Bajardi et al. [3] utilized mobile phone data to analyse the spatial spread of infectious diseases. By combining human mobility patterns with disease transmission dynamics, their research provided valuable insights into the impact of travel on the geographical dissemination of infections. Such studies bridge the gap between theoretical models and real-world scenarios, enhancing our understanding of epidemic dynamics in urban environments.

Similar studies regarding influence propagation from Watts and Strogatz (1998) [4] introduced the concept of 'small world' networks. This research aimed to reconcile two properties of networks which are usually viewed to be contradictory: a short average path length and a high degree of local clustering. The developed models which demonstrated the existence of both properties within the same networks, thus naming them 'small-world' networks. The small average path length along with the high degree of clustering proved to be a very efficient way of information transfer within a network. Short paths enabled quick transportation, while high clustering degree ensured higher rates of spread of information.

3. METHODOLOGY

3.1 Data Set Description

We have employed a synthetic social network dataset generated using the *Erdős-Rényi* model, a classical random graph model widely used for network analysis. The *Erdős-Rényi* model creates networks by connecting nodes with a specified probability, leading to a diverse range of network structures, from sparse to densely connected. A deeper understanding of the *Erdős-Rényi* model follows.

The *Erdős-Rényi* model was developed by mathematicians Paul Erdős and Alfréd Rényi to explore random graph structures. It starts from a set of nodes/vertices, where the edges between the nodes are created based on a parameter called the probability distribution. It is a function which describes the possibility of an event occurring. In

this case, the probability distribution will determine the probability of the formation of an edge between two successive nodes/vertices.

Network Generation Process:

Node Creation: We started by creating a set of nodes representing individuals within the social network. The total number of nodes, denoted as N, was determined based on the desired scale of the simulation. Each node was assigned a unique identifier, ranging from 0 to $N-1$.

Edge Formation: For each pair of nodes i and j $(\neq i = j)$, an edge between them was established with a probability p. This probability p determined the likelihood of a social connection between any two individuals. The resulting graph was undirected, meaning the relationship between node i and j was symmetric (if i was connected to j, then j was also connected to i).

By adjusting the values of N and p, we could create social networks of varying sizes and densities. A larger N led to larger networks, while changing p influenced the overall connectivity within the network. For low values of p, the resulting network tends to be sparse, with relatively few edges compared to the total possible connections. Nodes might have only a small number of connections, creating isolated clusters or components within the network. These parameters allowed us to explore the effects of network size and structure on the spread of infections. Such networks are called sparse networks. Higher values of p, on the other hand, lead to denser networks with a larger number of edges. Nodes are more interconnected, resulting in a higher average degree (the average number of connections per node) and a smaller average path length between nodes. These connections are known as dense networks.

The *Erdős-Rényi* model has been extensively used in theoretical studies. It serves as a tool for understanding network properties and the random graph theory. The model serves as a mode of simulation and analysis for evaluating network properties in complex systems, such as social networks, communication networks and biological networks

3.2 Models

In this study, we utilized the Independent Cascade Model (ICM) to simulate the spread of infections in the generated social network. The Independent Cascade Model is a widely employed probabilistic model for information diffusion in social networks. It captures the stochastic

nature of how information, innovations, or diseases spread through a network of interconnected individuals.

Independent Cascade Model:

In the Independent Cascade Model, the spread of the infection is characterized by the following key components:

Infection Probability (p): Each edge in the network is associated with a probability p representing the likelihood of the infection spreading from an infected node to its neighbouring susceptible node.

Infection Process: When an infected node attempts to infect its susceptible neighbours, the infection succeeds with a probability p. If successful, the neighbouring node becomes infected.

Probabilistic Spread: The spread of infection occurs probabilistically in discrete time steps. At each time step, infected nodes attempt to infect their susceptible neighbours independently with the probability p.

Mathematical Equation:

The probability $P(v \text{ becomes infected})$ that a susceptible node v becomes infected can be calculated using the following equation:

$$P(v \text{ becomes infected}) = 1-(1-p)^{\text{number of infected neighbours}}$$

This equation calculates the probability that node v becomes infected based on the probabilities of its infected neighbours attempting to infect v.

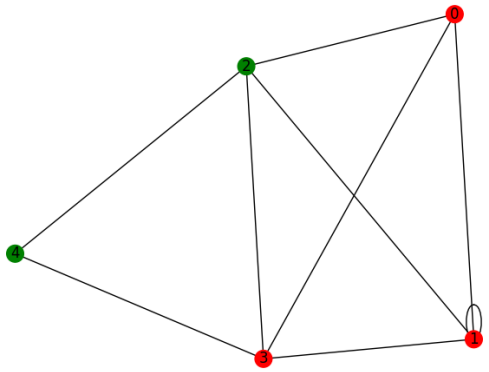
3.3 Evaluation

The model was tested multiple times with different sets of values in each trial to study the functioning of the model, as well to understand the spread of infection from one node to another through the network.

Below is one set of readings, where we ran simulations for a graph of 5 nodes with only 1 seed node.

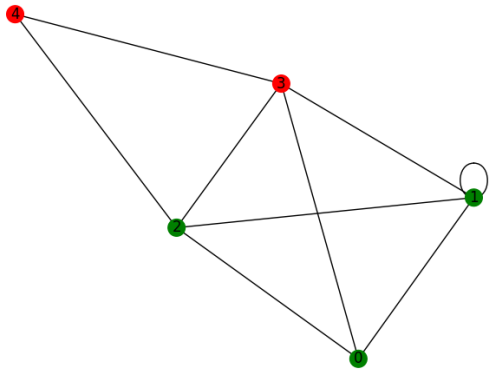
Trial	Infection Probability	Data Loss
1	0.1	60%
2	0.2	40%
3	0.3	0%

```
Enter the number of nodes in the graph (up to 100): 5
Enter adjacency matrix for the graph (1 for edge, 0 for no edge):
0 1 1 1 0
0 1 0 1 0
1 1 0 1 1
1 0 1 0 1
0 0 1 0 0
Enter the number of seed nodes (up to 5): 1
Enter the ID of seed node 1 (0 to 4): 2
Enter the infection probability (0 to 1): 0.1
Total Nodes: 5
Infected Nodes: 2
Data Loss Percentage: 60.00%
```



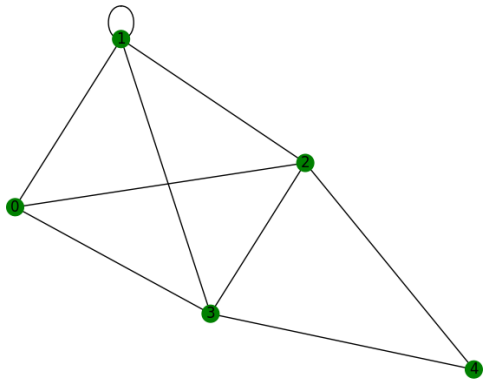
Trial 1

```
Enter the number of nodes in the graph (up to 100): 5
Enter adjacency matrix for the graph (1 for edge, 0 for no edge):
0 1 1 1 0
0 1 0 1 0
1 1 0 1 1
1 0 1 0 1
0 0 1 0 0
Enter the number of seed nodes (up to 5): 1
Enter the ID of seed node 1 (0 to 4): 2
Enter the infection probability (0 to 1): 0.2
Total Nodes: 5
Infected Nodes: 3
Data Loss Percentage: 40.00%
```



Trial 2

```
Enter the number of nodes in the graph (up to 100): 5
Enter adjacency matrix for the graph (1 for edge, 0 for no edge):
0 1 1 1 0
0 1 0 1 0
1 1 0 1 1
1 0 1 0 1
0 0 1 0 0
Enter the number of seed nodes (up to 5): 1
Enter the ID of seed node 1 (0 to 4): 2
Enter the infection probability (0 to 1): 0.3
Total Nodes: 5
Infected Nodes: 5
Data Loss Percentage: 0.00%
```



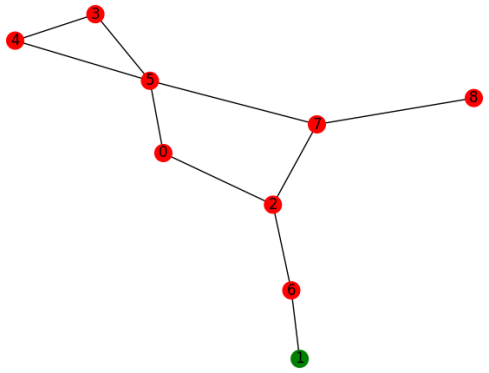
Trial 3

Here we have the trial results of the graph made of 5 nodes (as shown in the figures) and only 1 seed node. The seed node in this case is Node 2. The results show the increase in data transfer or the reduction in data loss with an increase in the infection probability.

We ran more simulations, this time with a different and more complex graph of 9 nodes, with 1 seed node. This was to demonstrate the impact of nodes with more connections within the network, and their role in data transfer.

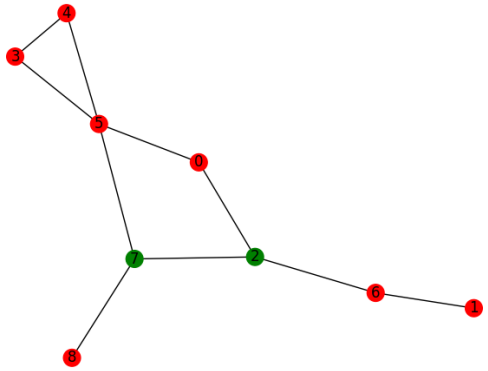
Trial	Node ID	Data Loss
1	1	88.89%
2	2	77.78%
3	5	22.22%

```
Enter the number of nodes in the graph (up to 100): 9
Enter adjacency matrix for the graph (1 for edge, 0 for no edge):
0 0 1 0 0 1 0 0 0
0 0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 1 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 0 1 0 0 0
0 0 0 1 0 0 0 1 0
0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0
Enter the number of seed nodes (up to 9): 1
Enter the ID of seed node 1 (0 to 8): 1
Enter the infection probability (0 to 1): 0.1
Total Nodes: 9
Infected Nodes: 1
Data Loss Percentage: 88.89%
```



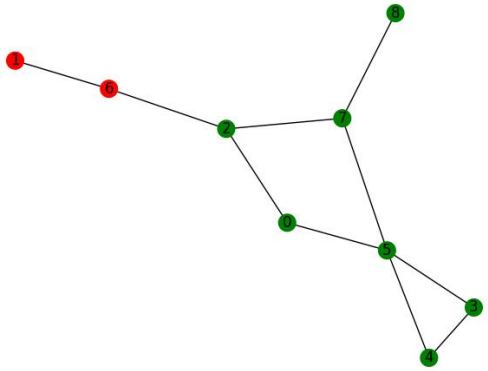
Trial 1

```
Enter the number of nodes in the graph (up to 100): 9
Enter adjacency matrix for the graph (1 for edge, 0 for no edge):
0 0 1 0 0 1 0 0 0
0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 1 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 0 1 0 0 0
0 0 0 1 0 0 0 1 0
0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0
Enter the number of seed nodes (up to 9): 1
Enter the ID of seed node 1 (0 to 8): 2
Enter the infection probability (0 to 1): 0.2
Total Nodes: 9
Infected Nodes: 2
Data Loss Percentage: 77.78%
```



Trial 2

```
Enter the number of nodes in the graph (up to 100): 9
Enter adjacency matrix for the graph (1 for edge, 0 for no edge):
0 0 1 0 0 1 0 0 0
0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 1 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 0 1 0 0 0
0 0 0 1 0 0 0 1 0
0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0
Enter the number of seed nodes (up to 9): 1
Enter the ID of seed node 1 (0 to 8): 5
Enter the infection probability (0 to 1): 0.2
Total Nodes: 9
Infected Nodes: 7
Data Loss Percentage: 22.22%
```



Trial 3

To demonstrate the importance of the location of the target node, we fixed the number of nodes and branches in the system. Also, the infection probability was fixed at 0.2. It is quite evident that the nodes which are situated towards the centre of the network tend to have more influence in the transfer of data. These 'influential' nodes are defined by the number of edges connected to them. In trial 1, the seed node is Node 1, which lies in the outskirts of the network, with only 1 edge. The low infection probability restricts the flow of the infection from Node 1 to Node 6, which in turn, completely inhibits the flow of infection. In trial 2, the seed node is set to Node 2, which has 3 edges to Node 0, Node 6 and Node 7. In this case, the transfer of infection is restricted from Node 2 to Node 0 and Node 6 because of the low infection probability. However, the infection spread is possible in the case of Node 2 to Node 7. The third trial has Node 5 as the seed node. It has the most edges within the graph – 4 in number: Node 5 to Node 0, Node 5 to Node 3, Node 5 to Node 4, and Node 5 to Node 7. The results showed that 7 out of 9 nodes were infected even with a small infection probability of 0.2. All the nodes in direct contact with Node 5 were infected, along with Node 2 and Node 8, which were not in direct contact with Node 5. In the case of Node 8, the infection must have travelled from Node 5 to Node 7 to Node 8. The infection of Node 2 could have arisen from a similar

propagation from Node 5 to Node 7 and to Node 2, or also from Node 5 to Node 0 to Node 2. From these readings we can clearly observe the degree of influence which is possessed by the nodes in the central region of the network, or nodes which have a higher count of edges.

4. RESULTS

Our analysis revealed notable insights into the infection spread dynamics. The infection spread rate varied based on the initial seed nodes, network connectivity, and infection probability. Nodes with minimal edges had a lesser probability to transfer the infection to its neighbouring nodes. Nodes which were situated in more centralised locations tend to control the propagation of the infection to a greater degree when compared to those located at the fringes/periphery of the network. Looking through the views of epidemiology, we can learn from our observations that a small number of initial infections could lead to a widespread epidemic, especially when targeting highly influential nodes with a large number of connections. The denser the network, the quicker the infection spreads across the nodes.

These results can help us to understand the behaviour of such networks. Our understanding will be essential for developing effective containment strategies for the same. One such method to supress/cure infections is targeted immunization of highly influential nodes. By identifying those nodes, minimal time and resources will be spent while giving a major impact to the network.

5. CONCLUSION

While our results shed light on various aspects of infection spread, it is crucial to acknowledge the limitations of our study. The simplified nature of the Independent Cascade Model does not capture the complexities of real-world interactions, including human behaviour, mobility patterns, and heterogeneous contact rates. Future research should explore more sophisticated models that consider these factors to enhance the accuracy of epidemic predictions.

In conclusion, future enhancements in the analysis of influence propagation and data loss in social networks should prioritize dynamic network models, content-based and behavioural analyses, privacy-preserving measures, multi-modal data sources, and advanced machine learning techniques. Incorporating these elements will lead to a more comprehensive understanding of the complex dynamics within social networks, enabling the development of robust strategies for predicting

influence propagation and mitigating the risks associated with data loss. Moreover, a user-centric approach, cross-platform analysis, robustness assessments, and a commitment to ethical considerations are essential components to ensure responsible and meaningful research practices in this evolving field. By addressing these aspects, researchers can contribute valuable insights that not only advance the theoretical understanding of social network dynamics but also inform practical measures for safeguarding user privacy and network integrity.

6. REFERENCES

[1] Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. Physical Review Letters, 86(14), 3200.

[2] Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 137-146).

[3] Bajardi, P., Poletto, C., Ramasco, J. J., Tizzoni, M., Colizza, V., & Vespignani, A. (2011). Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. PLoS ONE, 6(1), e16591.

[4] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440-442.

https://visiblenetworklabs.com/2021/04/16/understanding-network-centrality/

https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/#:~:text=Definition%3A%20Degree%20centrality%20assigns%20an,other%20nodes%20in%20the%20network.

https://dgarcia-eu.github.io

https://www.geeksforgeeks.org/degree-centrality-centrality-measure/