Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/79/2025.20135

Application of Graph Theory in Social Network Analysis

Shuo Wang[1,2,a,*]

[1]*University of California, Santa Barbara, Santa Barbara, USA*
[2]*Department of Mathematics, University of California, Santa Barbara, Santa Barbara, USA*
a. shuowang@ucsb.edu
**corresponding author*

Abstract: This literature review aims to discuss the application of graph theory in analyzing social and information networks. We first introduce some key network properties, such as clustering coefficient (transitivity), centrality, and diameter, which are crucial for understanding the dynamics of information dissemination within networks. Then, we talk about, based on these properties, how graph theory can be utilized to analyze social and information networks. Lastly, we provide an overview of various fundamental social and information models including the SIR model and the Linear Threshold model. For the SIR model, we go over the definition of the SIR model, explore the mathematical methods applied to analyze the SIR model, provide one example of how the SIR model reflects the nodes' status in the network, and discuss the application of the SIR model in Covid-19 epidemic. For the Linear Threshold model, we review its basic properties and explain how it can be used to calculate the maximum influence in a network.

Keywords: Graph Theory, Social and Information Networks, SIR model, and Linear Threshold model.

1. Introduction

In modern society, the rapid development of the internet and online platforms has transformed the way information is exchanged and disseminated. Understanding the structure, dynamics, and mechanisms behind information spreading within society has become increasingly important. Under such circumstances, social and information networks, serving as reflections of exchanges of information, social influences, and ideas between people, organizations, and groups, are crucial for exploring the information spreading. These networks offer valuable insights into how information propagates and how it can be shaped by various factors. In order to quantitatively analyze the information spreading in these networks and find out the factors influencing the spreading of information, statistical methods are indispensable for an effective and quick analysis. By employing graph theory and related techniques, researchers can gain deeper insights into the fundamental characteristics of social and information networks. For example, the classical sociological theories always believe the configuration of the network and the connections between nodes can reveal the relationship between personal health and community resilience, which is important in fields like economics, public health, and criminology [1] [2]. Usually, researchers use quantitative or qualitative statistical methods to analyze and collect information, which allows them to understand key terms of data like density and strength. However, due to the complexity of the real-world example and the

DOI: 10.54254/2753-8818/79/2025.20135

huge workload of the statistical methods, it is hard for researchers to realize some properties of the information, such as the clustering effects, and analyze a large amount of information in an efficient way [3]. On the other hand, by creating suitable and desired graphs, researchers can use the characteristics of the networks to learn about the spreading of the information and the overall distribution of the nodes like clustering coefficient and centrality. Under such circumstances, scientists can analyze the behaviors of information in social structures and facilitate the creation and application of computation models [4]. In this review, we will learn about some important graph theory parameters that are useful in the analysis of social and information networks and review some fundamental social and information models.

2. Crucial Parameters of Networks in Graph Theory

Networks usually refer to a graph consisting of multiple nodes or vertices connected by edges. In social and information networks, the whole graph can be considered an ensemble of information such that nodes represent the sources and edges represent the connection between these sources. In graph theory, multiple properties of networks can facilitate researchers in analyzing and learning the details and distribution of the whole network. Hence, researchers can use the properties of networks to understand the structures and dynamics of information and try to control them. In this chapter, we will clarify the definition and meaning of transitivity (clustering coefficient), centrality, and diameter.

2.1. Clustering Coefficient (Transitivity)

The clustering coefficient is defined as the degree that the nodes in a graph tend to cluster together and on the other hand, it represents the fraction of triangles in the network. To be more specific, there are two types of clustering coefficients: local clustering coefficient and global clustering coefficient. The local clustering coefficient measures the connection of a node's neighborhood, while the global clustering coefficient represents the ratio of existing edges and the possible connections between nodes, so the global clustering coefficient can demonstrate the overall clustering in the whole network [5]. The concepts of these two clustering coefficients allow researchers to understand the interconnected subgroups of the entire network and help to identify the structure of the networks, helping improve the understanding of some structural properties of complex networks, such as social networks, biological networks, and technological networks [6].

2.2. Centrality

Centrality measures the ranking of nodes, so it can be used to determine the most important and influential node in a network [6]. Based on the definitions of importance or type of influence of the network, there are various types of centralities, including betweenness centrality and closeness centrality. Betweenness Centrality refers to the fraction of the shortest path passing through a node, it enumerates the number of times a special node passes by the shortest paths between two other nodes. Hence, the node with high betweenness centrality is crucial for highlighting the information flow in the network [7]. On the other hand, Closeness centrality is defined as the inverse of the average distance of a node to all other nodes in the network. Therefore, the closeness centralities of a network indicate the overall distribution of a graph, especially the compactness of a network. However, closeness centrality might be indispensable when the degree distribution is known, since researchers have demonstrated that the inverse of the closeness centrality is relatively proportional to the logarithm of the degree number [8].

DOI: 10.54254/2753-8818/79/2025.20135

2.3. Diameter

Diameter in the graph theory refers to the largest distance between two connected nodes. The diameter is the key to depicting the overall structure and connectivity of a network, so it has been applied in a large number of social and information networks. One of the most influential ideas related to diameters is the six-degree separation. This theory proposes that in the real world, any two arbitrary people can be connected in a social chain with a length of at most six. Past research revealed that for a random generalized network, the typical separation between two nodes in a graph is close to the fraction between the logarithm of the total number of nodes in the network and the logarithm of average edges per node [5]. If we consider the whole society as a social network, the total number of nodes is approximately 7,000,000,000, and the average edges per node are approximately 50. Based on this data, the calculated typical node distance is 5.79, which is less than 6. This result indicates that the world can be considered a “small world” network. Even though the network size is large enough, the diameter of the network is still small.

3. Social and Information Network Models

The properties of social and information networks reveal that they can be used effectively in many real-life examples. For example, a social and information network can be applied in epidemiology while it is regarded as a virus propagation information assembling like the propagation of Covid-19 [9]. In this case, nodes represent individuals in the network and information spreading represents the infection of the individual or recovery of an individual. However, in real life, many factors can affect information propagation. As a result, it is impossible to design a precise social and information network that is able to deal with such imponderable information. So, developing a specific social and information model to simulate the instance in real life is necessary. Researchers have divided the social and information into two different parts based on the information diffusion process, methods, and influence, and created two models accordingly: the explanatory model and the predictive model. Explanatory models describe how information propagation within a network, can also be divided into two types of models: epidemic models and influence models. Epidemic models as the name says always are used to simulate the propagation of an epidemic. Influence models on the other hand focus on the influence of some specific nodes in the network. Predictive models describe the prediction of information propagation in a network, and they include the independent cascade model (ICM), linear threshold model (LTM), and game theory model (GTM). The ICM model focuses on activating inactive nodes in social networks and the node can only be active once while the nodes can be active multiple times only if the influence on that node exceeds a certain level. The GTM model is used to consider the methods that maximize the profits of spreading information [9]. In this review, we will mainly discuss an epidemic model SIR and a linear threshold model.

3.1. SIR model

SIR model stands for susceptible (S), infectious (I), and recovered (R). These three terms refer to three stages in viral spreading: susceptible means the node in the network has no awareness of the information, infectious means the node does know the information and is spreading the information right now, and recovered means the node no longer spreads the information anymore. In a fully connected network, mathematical methods are required to analyze the SIR model. We use β to represent the average contact rate of the nodes in a network which means the average possibility that a person is in contact per unit time. γ is defined as the recovery rate which is the probability that an infectious node becomes recovered in the next unit of time. To find out the overall distribution of the network, $s(t)$ is defined as the fraction of susceptible nodes at time t, $x(t)$ is defined as the fraction of

DOI: 10.54254/2753-8818/79/2025.20135

infected nodes at time t, and $r(t)$ is defined as the fraction of recovered time t in the network. In such case, we can use parameters to estimate the rate of change of SIR states:

$$\begin{aligned} \frac{ds}{dt} &= -\beta \dot{s}\dot{x} \\ \frac{dx}{dt} &= \beta \dot{s}\dot{x} - x\dot{\gamma} \\ \frac{dr}{dt} &= x\dot{\gamma} \end{aligned} \tag{1}$$

The solutions to these equations are:

$$\begin{aligned} s &= s_0 e^{-\frac{\beta}{\gamma} r} \\ \frac{dr}{dt} &= \gamma\left(1 - r - s_0 e^{-\frac{\beta}{\gamma} r}\right) \\ s + x + r &= 1 \end{aligned} \tag{2}$$

By solving the problem numerically, we can finally get the solution of s, x, and r. Therefore, scientists are able to analyze the fraction of the population as a function of time t [10]. Figure 1 is an example of application of SIR model in estimating the fraction of susceptible, infected, and recovered people in a network as the increase of time. Under a certain initial condition, the results of the susceptible, infected, and recovered rate are also fit with the corresponding solutions of s, x, and r.

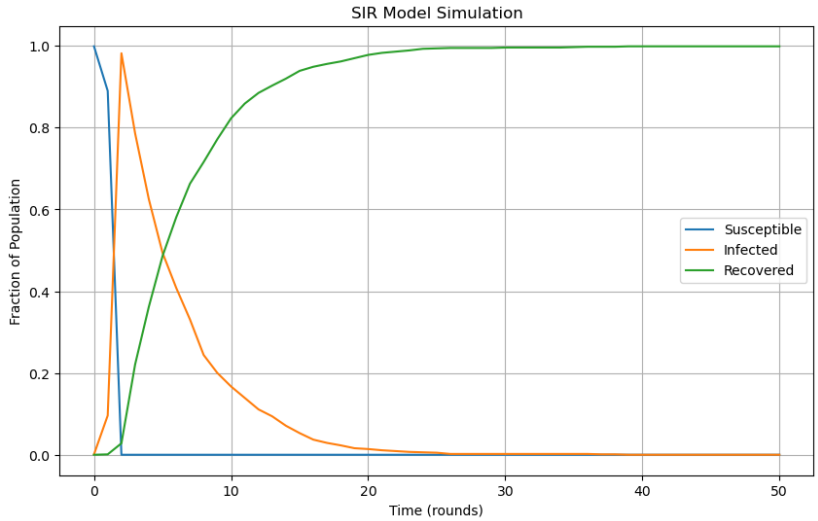

Figure 1: Population of susceptible, infected and recovered individuals relative to time t in a standard SIR model. x-axis represents time and y-axis represents population fraction.

Due to the characteristics of the node's representations, the SIR model is usually applied in the field of virus spreading and it is one of the most commonly used models in epidemiology. During the pandemic of Covid-19, researchers have applied the SIR models to estimate the progression of the pandemic using the existing Covid-19 data and they realized that compared to many more complicated models, the SIR model can effectively and simply predict the epidemic dynamics using fewer requirements [11]. However, the SIR model is a simple model with several limitations. It cannot handle the problem with more restrictions or account complex interaction between nodes and connections. Hence, the exploration of new models and more complicated models is inevitable.

DOI: 10.54254/2753-8818/79/2025.20135

3.2. Linear Threshold Model (LTM)

In the Linear Threshold Model, the nodes have two states: active and inactive, and each node has a threshold $\boldsymbol{\tau}$ such that the node can only be active if and only if the number of active neighbor node m and the number of total neighbor nodes k satisfy the condition $\frac{\boldsymbol{m}}{\boldsymbol{k}} \geq \boldsymbol{\tau}$ [12]. Once the node is activated, the node is no longer be able to deactivate, so, the nodes that have potential to be activated will eventually become active if there is no time limit and there is a possibility that a spreading event in linear threshold model evolves into a global cascade like contagious epidemic. To learn about the threshold of global cascade, researchers define the probability that a node with degree k which means k neighbors turns active with one active connection be $p_k = \mathrm{P}(\tau \leq \frac{1}{k})$. Under such condition, the threshold for global cascade will be:

$$\mathrm{T} = \frac{\langle \mathrm{p_k}\left(\mathrm{k}^2 - \mathrm{k}\right)\rangle}{\langle \mathrm{k}\rangle} \tag{3}$$

The equation 3.3, to be more specific, represents the expected number of new activations that a node which just turn active will generate. If this value is large than 1, there will be global cascades while if the value is less than or equal to 1, there will be small cascades. Therefore, for situations under the linear threshold model, scientists can calculate the maximization influence in large networks [13].

4. Discussion

In this literature review, we discuss a comprehensive foundation of key terms graph theory and their applications in social and information networks. They have multiple advantages in the analysis of the information propagation, but they also have some limitations in their practical use.

The review of some important parameters in graph theory including clustering coefficient (transitivity), centrality, and diameter uses some references that introduce the definition of these parameters and the insights into how these parameters reflect the network structure, node connections, and the influence of some specific nodes. These references always present a well-defined definition of the parameter and reveal part of the application of these parameters. However, these references always focus on static models and ignore their application in temporal dynamics, in which the network will evolve with time. Some recent theories have been incorporated into the fields of temporal or dynamics networks, and this extends the scope of the network application and enhances the development of some new network theories. For example, researchers have developed a dynamic-sensitive centrality measure that can locate the local influential nodes in temporal networks, which captures the dynamic change of the network over time. This approach enables researchers to effectively identify the key influencers in networks and outperforms traditional centralities in accuracy in temporal networks [14].

The SIR model can be applied to many real-life situations and offers a practical framework for understanding the spread of infections in social networks [10]. During the COVID-19 epidemic, researchers have shown that the SIR model can estimate the spreading of COVID-19 effectively even if it is simply [11]. Hence, the simplicity and effectiveness of the model make it a valuable tool for epidemiological analysis, allowing for rapid assessments of disease spread. However, the SIR model's limitations lie in its assumptions of homogeneity among nodes, as it does not capture complex interactions such as super-spreader events or varied recovery rates among populations. Enhancing the model with features like multi-layer structures or time-varying connections could extend the situations in that it can be used, but the structure of the model still makes it unable to handle cases in which researchers pay attention to behaviors spreading.

DOI: 10.54254/2753-8818/79/2025.20135

The Linear Threshold and Independent Cascade models unlike the SIR model mainly stimulate the changes of behaviors or propagation of actions. These models are important in exploring how initial influences spread across the networks, especially because they illustrate how thresholds and probabilistic interactions can drive large-scale behavior change. The improvement of the multi-layer network model can increase the application of the model in real life and handle more complicated situations. However, these models have limitations, since they assume static influence thresholds and do not account for complex social interactions, feedback loops, or adaptive behaviors among nodes, which can reduce their accuracy in dynamic networks.

5. Conclusion

In this review, we study the importance of graph theory in social and information networks, discussing how key parameters in graph theory and models help to understand the network structure and predict the information propagation within the network. To be more specific, we talk about the fundamental network properties like clustering coefficient, centrality, and diameter and how they can be applied in estimating nodes' influence in the network, connectivity of the nodes, and the overall structure of the network. In addition, we explore the SIR model and the Linear Threshold model and how they can be applied in networks. The SIR model can be used as a tool for understanding the spreading of infected virus while the Linear Threshold model can be used to capture the behavioral changes or influence cascade within the network. Even though, for each parameter and model, there are underlying limitations, there are still many aspects that we can apply in the social and information networks. Furthermore, these limitations indicate the potential for the field to be explored and created. There has been research focusing on how multi-layer models or temporal networks reflect real-life networks. For example, there is research modeling the SIR in multi-layer networks and showing the impact of the epidemic in multi-layer systems [15]. Overall, graph theory provides a powerful tool for scientists to analyze social and information networks, and the sinuousness of real-life cases makes it possible for researchers to design more complicated and elaborate models that can be applied in reality.

References

[1] *Granovetter, M.S. (1973) The Strength of Weak Ties. American Journal of Sociology, 78, 1360-1380.*
[2] *Christakis, N.A. and Fowler, J.H. (2007) The Spread of Obesity in a Large Social Network over 32 Years. The New England Journal of Medicine, 357, 370-379.*
[3] *Newman, M.E.J. (2003) The Structure and Function of Complex Networks. SIAM Review, 45, 167-256.*
[4] *Pastor-Satorras, R., & Vespignani, A. (2001) Epidemic Spreading in Scale-free Networks. Physical Review Letters 86, 3200-3203.*
[5] *Watts, D. and Strogatz, S. (1998) Collective dynamics of 'small-world' networks. Nature 393, 440–442.*
[6] *Newman, M. E. J. (2010) Networks: An Introduction. Oxford University Press.*
[7] *Freeman, L. C. (1977) A set of measures of centrality based on betweenness. Sociometry 40, 35-41.*
[8] *Evans, T.S. and Chen, B. (2022) Linking the network centrality measures closeness and degree. Commun Phys 5, 172.*
[9] *Li M., Wang X., Gao K. and Zhang S. (2017) A survey on information diffusion in online social networks: models and methods. Information 8, 118.*
[10] *Kermack W.O. and McKendrick A.G., (1927) A contribution to the mathematical theory of epidemics. The Royal Society 115, 700–721.*
[11] *Postnikov, E.B. (2020) "Estimation of COVID-19 dynamics "on a back-of-envelope": Does the simplest SIR model provide quantitative parameters and predictions?". Chaos, solitons, and fractals vol. 135, 109841.*
[12] *Watts D.J., (2002) A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. U.S.A.99, 5766-5771.*
[13] *Chen W., Yuan Y. and Zhang L., (2010) "Scalable Influence Maximization in Social Networks under the Linear Threshold Model," 2010 IEEE International Conference on Data Mining, Sydney, NSW, Australia, 88-97.*
[14] *Huang, DW. and Yu, ZG. (2017) Dynamic-Sensitive centrality of nodes in temporal networks. Sci Rep7, 41454.*

DOI: 10.54254/2753-8818/79/2025.20135

[15] Turker, M., Bingol, H.O. [2023] Multi-layer network approach in modeling epidemics in an urban town. Eur. Phys. J. B 96, 16.