# Second-Order Random Walk in Complex Networks

Zongheng Gu
NO.1 Middle School affiliated to Central China Normal University, Wuhan, Hubei 430223, China
Wisadel0621@outlook.com

**Abstract.** In the era of complex networks, it is important to understand the network structure and user behavior. Random walks, especially second-order random walks, provide important tools for analyzing and optimizing networks. This paper combines second-order random walks with graph theory to address the challenges in various networks. The principle, method, and application are expounded in detail through theoretical analysis and practical cases. The research identifies key nodes and discovers communities in social and academic networks to provide a scientific basis for information management and precision marketing. However, challenges such as computational pressure and parameter selection remain. Future solutions include leveraging distributed computing, building theoretical frameworks, and using machine learning to optimize parameters and improve data quality. Overall, this study provides a new perspective for complex network analysis and lays a foundation for further development. Second-order random walk has been effective in many fields, and it is expected to promote intelligent development in more complex network scenarios in the future.

**Keywords:** Second-Order Random Walk, Graph Theory, Complex Network, Community Detection, Key Node Identification.

## 1. Introduction

The Internet is an integral part of today's society. In the social network, understanding the relationship between users and the law of user information dissemination is beneficial. It enables more accurate personalized content to be pushed to each customer, thereby enhancing the Internet experience of users. In the urban traffic network, exploring the changes in traffic flow and the connection relationships between roads is the key to optimizing the allocation of resources and alleviating traffic congestion. However, when faced with such complex network problems, traditional analysis methods often find it difficult to comprehensively and effectively address these challenges.

Based on this, random walks have been widely used in various network analysis and optimization problems. It is also called a stochastic process, and it describes a path that is a series of random steps in mathematical space, which is widely used in network models. Moreover, strongly connected digraphs, as an effective network model, are especially suitable for describing those networks where any two nodes can reach each other through a bidirectional path. These complex networks can then be simplified into strongly connected digraphs. Therefore, strongly connected digraphs have important application value in social networks, traffic flow analysis, and other fields.

There are different sampling methods of random walk in strongly connected digraphs. When selecting the next sample, first-order random sampling only depends on the currently selected sample itself, which makes the first-order random walk unable to remember the characteristics of the model in the long term, and the prediction effect is often ineffective when unexpected events occur. On the other hand, the second-order random walk not only depends on the current sample when selecting the next sample but also depends on the samples selected the previous or multiple times, which is capable of perfectly coping with the prediction of unexpected situations. Therefore, the selection probability of second-order random sampling on a strongly connected digraph can be regarded as the selection probability of random sampling on its dual graph.

In this regard, there has been a lot of research that has made some progress. For example, Lu used the harmonic mean to improve the accuracy of estimating the average degree of Optical Switch Network (OSN) [1]. Rao and his team provide a more efficient tool for complex transportation network work, namely First-Order Gradient Supervision (FOGS) [2]. In the study of community discovery, the Approximate Personalized Page Rank (APPR) method uses standardization to sort the APPR vector nodes and employs a cleaning technique to determine the target community. In addition, the Multi-Walker Chain (MWC) model changes the starting vector to focus on the target community [3]. Zhang proposes a New Clustering Hierarchical Random Walk (NCHRW), a sampling algorithm based on the clustering coefficients of node neighborhoods, and combines the concept of set conductance with that of clustering Hierarchical Random Walk. Considering the current node degree, the number of adjacent public neighbor nodes, and the clustering coefficient of neighboring nodes, the selection skip selection is optimized [4]. Although these methods provide powerful tools for solving multiple problems in complex networks, there are still some challenges and room for optimization.

In this paper, the theory and advantages of the method of combining theoretical analysis with real life are elaborated meticulously. This paper aims to reveal its application value, analyze both its merits and drawbacks, and improve the previous studies.

## 2. Basic Concepts and Principles

### 2.1. Basic Concepts of Graph Theory

The graph G=(V(G), E(G)), where V stands for the set of vertices and E stands for the set of edges. The directed graph indicates that all edges in the graph G are directional, and (u, v) is usually used to indicate the edge of the vertex u pointing to the vertex v. Correspondingly, an undirected graph G means that no edge in the graph G is directional.

Graphs are divided into connected graphs, strongly connected graphs, weakly connected graphs, and unconnected graphs. Connected graphs usually appear only in undirected graphs. If any two points in an undirected graph have a path that connects them, then this undirected graph is defined as a connected undirected graph. Weakly connected graphs and strongly connected graphs are concepts in directed graphs. For a weakly connected digraph, when all its directed edges are regarded as undirected edges and meet the requirements of a connected undirected graph, it is considered a weakly connected digraph. The definition of a strongly connected digraph is that from any point in the digraph, there is a path to any other point on the graph.

### 2.2. The Basic Concept of a Random Walk and its Relation to Markov Chains

A regular simple random walk, also known as a first-order random walk, typically assumes that under discrete time steps, a particle navigates within one-dimensional or higher-dimensional space. Each step of its movement is dictated by a probability distribution, encompassing both the direction and the distance.

The disparity between a second-order random walk and a first-order random walk lies in the fact that the alteration at each moment hinges not only on the state of the preceding moment but also on the prior time instants.

Markov chain is a stochastic process characterized by Markov properties. The Markov property dictates that, given the knowledge of the current state, the future state is solely related to the current one. In this regard, the first-order random walk can be seen as an instantiation of the Markov property.

However, the second-order random walk, due to its state transition relying on the previous two moments, fails to conform to the Markov property. It differentiates itself from the traditional Markov chain and can be regarded as a higher-order Markov chain as it incorporates more remote state information. Similar to an extended variant of the Markov chain, it accounts for the influence of additional historical states in the state transition probabilities.

### 2.3. Principle and Method

The key nodes are identified based on the combination of second-order random walk and graph theory, and the direct and indirect influences of nodes are considered comprehensively. The influence score of nodes Is is calculated by second-order random walk, where Is can be formulated as

Is=$\sum_i p_i \times w_i$($p_i$ represents the access probability of node i, $w_i$ represents the corresponding weight factor). The influence of nodes is further measured by the access probability and the control ability of information transmission (such as the centrality of interlocution). At the same time, combined with the centrality index in graph theory (such as degree centrality, proximity centrality, etc.), the node is evaluated comprehensively. Nodes with high impact scores and high centrality are identified as critical nodes. In the specific calculation process, firstly, the dynamic influence index of nodes is calculated according to the second-order random walk model, and then weighted summation is performed with the graph theory centrality index to obtain the final node importance score, and the key nodes are determined according to the score ranking.

In terms of community discovery, nodes are more likely to move within the community in the second-order random walk process, and can also find weak connections across community boundaries. The community to which the node belongs can be determined by statistical indicators such as the visit frequency and residence time of the node during its migration. Specifically, nodes with a high frequency of mutual access are divided into the same community. For example, a second-order random walk starts with multiple initial nodes, and after a certain number of iterations, the similarity between the nodes is calculated (such as based on common access frequency). Then, the K-Means algorithm is applied. In the K-Means algorithm, the number of clusters k needs to be preset, and the nodes are randomly initialized as k centroids. Subsequently, each node is assigned to the nearest centroid based on a certain distance metric (e.g., Euclidean distance). The centroids are then updated iteratively until convergence, and finally, the nodes are divided into different communities.

## 3. Real Case Analysis

GAO conducted experiments on the user relationship data set of an online social platform provided by GAO, aiming to compare the performance of second-order random walk and traditional Girvan-Newman algorithm in community discovery [3].

In the process of the experiment, two methods were applied to the data set respectively. Moreover, the effect of the algorithm was measured by analyzing the node similarity within the community, the connection between the communities, and the modularity of the community division quality index [3]. Here, “modularity” is an important indicator that reflects the quality of community division, with higher modularity values typically indicating better-structured and more distinct communities [3].

The experimental results show that the second-order random walk method can find a finer community structure. Furthermore, the nodes within the community are more similar and the connections between the communities are clearer [3]. By evaluating the quality indicators of community division (such as modularity), the second-order random walk method can obtain a higher modularity value, indicating that its community division effect is better [3].

For example, in the data set of the social platform, the second-order random walk method identified some small communities hidden in the complex network structure, whose internal users had highly similar interests, but could not be accurately divided in the traditional method [3].

The social networks, provide a solid scientific basis for information dissemination management and intervention. Specifically, it can be used to timely curb the spread of rumors and guide the effective dissemination of correct information.

In the context of social networks investigated by WANG and his research team, within the analysis of academic cooperative networks (as opposed to complex networks like social networks), varying parameter settings can potentially yield diverse key node identification outcomes. Consequently, determining the optimal combination of parameters to precisely mirror the network structure and node relationships has emerged as an urgent issue demanding resolution. To address this, a methodology integrating second-order random walk and graph theory has been employed to identify crucial nodes [5].

Within this academic cooperation network, nodes symbolize scholars, while edges denote collaborative relationships among them [5]. The second-order random walk model is utilized to gauge the influence of scholars during the process of knowledge dissemination. Meanwhile, graph theory indices such as associative centrality and betweenness centrality are harnessed to single out those scholars who play a pivotal role in academic communication [5]. These key scholars not only possess a significant academic output (evinced by a high degree of centrality) but also fulfill a crucial function in bridging disparate academic groups and facilitating cross-domain knowledge dissemination (attested by a high betweenness centrality) [5]. This approach is capable of more comprehensively capturing the sway of scholars within the academic network and uncovering some individuals who, though overlooked in traditional analyses, in fact play an essential part [5].

## 4. Existing Problems and Challenges

First, second-order random walks are confronted with a tremendous computational burden when dealing with large-scale network data. As the scale of complex networks such as social networks and transportation networks perpetually expands, the number of nodes and edges skyrockets. For instance, in a colossal social network populated by a vast number of users, computing nodes' second-order neighbor information and performing random walk simulations predicated thereon entail a substantial demand for computing resources, especially in terms of CPU utilization and memory footprint [3]. It is the traditional computing devices and algorithms that often grapple with the scarcity of memory and protracted computing time when processing these data. This will grievously limit the application of this method in scenarios with stringent real-time requirements, like real-time traffic flow prediction or real-time information dissemination analysis on mammoth social platforms.

To enhance computational efficiency, compared with brute-force computation, current approximation algorithms like Metropolis-Hastings sampling or importance sampling technology present a more viable option, albeit with the drawback of potential accuracy loss. For example, during second-order random walk sampling, micro-cluster structures or weak ties that play a crucial role in network dynamics might be overlooked due to incomplete sampling, thus impinging on the final analysis result. This is an essential problem that necessitates careful consideration and resolution in practical applications.

Secondly, there is a dearth of systematic theoretical guidance for effective algorithm parameter tuning. To ascertain the weight distribution of second-order neighbors, the selection of walking steps, and the integration mode with graph theory indexes, it frequently has to resort to extensive experiments and heuristic approaches. Different types of networks possess distinct structural characteristics and application requirements. It is the distinctiveness of different network types that demands specialized algorithm optimization research. For example, there is a significant difference between the information transmission mode in social networks and the traffic distribution rule in traffic networks (refer to the research on relevant network structure and application characteristics). In some social networks, researchers have started to focus on leveraging user behavior data to optimize the weight of second-order neighbors, yet much remains to up.

However, the research in this area is still comparatively insufficient, and without further breakthroughs, it will continue to limit the effective application of this method in different fields. In the fields of smart grid management and epidemic spread prediction, the lack of optimized algorithms has hampered the full utilization of second-order random walks.

Thirdly, the parametric configuration in the second-order random walk model has a profound effect on the results. The model encompasses several parameters, such as the calculation parameters of transition probability, the weighting of second-order neighbor influence, etc. It is the small changes in these parameters, known as parameter perturbation, that can trigger profound deviations in the results. In practical applications, it is exceedingly challenging to ascertain the appropriate parameter values. Take social network analysis as an example, in the community discovery task, if the second-order neighbor influence weight is not set properly, the community division result may be inaccurate. At present, it mainly depends on experience and trial and error to select parameters, and there is no effective automatic selection method. This not only augments the difficulty of model application but also renders the performance of the model unstable in different data sets or application scenarios. For example, parameters optimized on one dataset may not yield the desired results when applied to another dataset with different structural characteristics.

In addition, although the method has a certain universality, it still needs to be refined in the face of different types and structures of networks. Compared with the ideal model that can adaptively capture network dynamics, the existing models fall short in handling highly dynamic social networks. For example, for highly dynamic social networks (such as instant messaging networks) or hierarchical networks (such as enterprise organizational structure networks), the existing models are difficult to fully capture their characteristics and cannot accurately reflect the key information and dynamic changes in the network, thus affecting the accuracy of analysis and decision-making. Recent research has proposed using deep learning techniques to automatically learn optimal parameters, but more exploration is needed.

Finally, the quality and integrity of data in practical applications have a crucial impact on the effectiveness of the combination of second-order random walk and graph theory. The acquired network data frequently presents diverse quality problems. For instance, in social networks, there may be fragmented user information, spurious accounts, and data integrity violations. In the community discovery task, it is precisely the fragmented user information that mainly leads to an inaccurate characterization of node features, subsequently affecting the accuracy of community division. In the prediction of information transmission, data loss will cause the model to be unable to accurately capture the initial conditions and transmission paths of the information flow, and thus reduce the reliability of prediction.

At the same time, the dynamic update of data also poses difficulties for the application of the model. The relationships and behaviors of users in social networks are in a constant state of flux, and the model must constantly adapt to new data. Given such ceaseless changes, the model needs to smoothly adapt to new data without augmenting the complexity of maintenance and update. Although some incremental learning techniques have been proposed to resolve this issue, they still face challenges in dealing with large-scale and rapid data alterations.

## 5. Future Solutions

In terms of computing pressure, it is with the help of distributed technologies such as cloud computing and cluster computing that the large-scale computing task can be deftly split into multiple nodes for parallel processing, which can fully leverage the cluster computing power and greatly reduce the computing time.

At the same time, academia and industry should collaborate to root in basic mathematical models, delve into the internal relations between second-order random walks and different network structures, and summarize the universal optimization principle through rigorous derivation. Establish groups dedicated to key networks like social networking and transportation, customize algorithm optimization schemes according to their characteristics, and then amass successful experiences to erect a cross-domain knowledge base, laying a solid theoretical foundation for subsequent research.

The parameter selection problem calls for machine learning, especially reinforcement learning to play a pivotal role. Provided that sufficient training data is available, a parameter optimization framework based on Deep Q-Network was constructed. By inputting network characteristics and task objectives and using community division module degrees and other indicators as reward feedback, the model can be driven to self-optimize. It can also build a parameter sensitivity analysis model to predict the impact of parameter changes in advance, assist in accurately locking key parameters, reduce trial and error costs, and stabilize model performance.

For data quality problems, first, strengthen pre-processing. Utilize data cleaning technology to weed out fake social network accounts and impute missing values, restoring the true and complete panorama of data with the help of multiple imputation and machine learning imputation algorithms. Introduce incremental learning to dynamic data, so that the model can track changes in real time and update itself dynamically. At the same time, construct a data quality monitoring system to issue timely warnings and conduct repairs when problems arise, ensuring the data quality for model operation. Through these schemes, the second-order random walk is expected to break through the dilemma and contribute to the development of the industry.

## 6. Conclusion

This paper centers on second-order random walks on strongly connected digraphs. The main part elaborates on its principle, methodology, and application. The results demonstrate that it exhibits remarkable performance in community discovery and key node identification within social and academic networks, which in turn facilitates precision marketing and traffic layout optimization.

These findings offer novel ideas for complex network analysis and extend the application of graph theory combined with random walk. However, this paper does possess certain limitations, such as the hurdle in processing large-scale data, the deficiency of theoretical optimization guidance, the complexity of parameter determination, and the impact of data quality. Future research should focus on surmounting these problems, enhancing computing efficiency, and exploring universal strategies for different networks, to enhance the applicability and robustness of this method in various complex network scenarios.

At present, the method has come to the fore in the field of social and transportation. For instance, social network marketing, has enabled companies to identify potential customer groups with higher precision, leading to a significant increase in conversion rates. In the future, with the improvement of technology and the continuous integration of artificial intelligence and the Internet of Things, it is expected that this method will be extensively utilized in more complex network scenarios and play an even more crucial role in optimizing intelligent transportation systems and facilitating personalized services in various industries.

## References

[1] Jianguo L, Dingding L. Sampling online social networks by a random walk. In Proceedings of the First ACM international workshop on hot topics on interdisciplinary social networks research, 2012, (8):33-40.

[2] Xuan R, Hao W, Liang Z, et al. FOGS: First-order gradient Supervision with Learning-based Graph for Traffic Flow Forecasting. In IJCAI, 2022, (7): 3926-3932.

[3] Yang G, Hongli Z. A review of community detection methods based on random walks. Journal of Communications, 2023, 44(06): 198-210.

[4] Mengyao Z. Research on community detection algorithm based on generative adversarial network. Chongqing University of Technology, 2023.

[5] Yuzhe W, Jinghua Y, Fanliang B, et al. RMFKAN: A network water army detection method based on improved graph Mamba. Computer Science and Technology, 1-18 [2024-12-24].