# Graph Generation and Diffusion using Random Walks

Wenyu Cai [1, *], Gilbert Chen Ye [2], Hao Zhou [3]

[1] East China Normal University, Shanghai, China

[2] University of Toronto, Toronto, Canada

[3] Nanjing University, Nanjing, China

* Corresponding Author Email: caiwenyuok@sina.com

**Abstract.** Simulations are often used in the study of metro network systems and the interactions of passengers with such systems in the real life. Graph theory is used to represent such metro systems. Simple random graphs are generated using a random graph generation algorithm revolving around random walks. The goal is to use such graphs to analyze the effects of the topologies of the graph parallel to the events which happen in real metro systems. This is done through random walks on the graphs by Monte Carlo Simulation of those random walkers. The simulations showed that the degree of a node in the graph has a near linear relationship with the number of times a specific node has been visited.

**Keywords:** Monte Carlo method, random walk, graph theory, graph generation, simple random graphs.

## 1. Introduction

Due to its affordable fares and high level of convenience, the metro is currently one of the primary forms of transportation in many modern cities all over the world. By 2021, there were metro systems in 205 cities in 61 nations around the world, including 43 in China. The Shanghai Metro is the largest metro system in the world, with a total length of 825 kilometers, a total of 20 lines, 407 stations, and an annual passenger volume of 3.57 billion in 2021 [1]. However, even such a large Shanghai metro is still incredibly crowded during the daily rush hour due to the enormous population base. Another situation where the metro system is put under a lot of stress is when a big event, like a ball game, finishes and tens of thousands of people rush into the closest station at almost the same moment. In such a problem that a significant number of people enter the metro network from a single station, the relationship between the structure of metro system and the degree of passenger flow diffusion needs to be understood to propose more efficient network structures.

Graph theory is often used to give mathematical representations of transportation networks, which more intuitively reflects the nature of transportation networks through graphical simplification [2]. Three categories can be used to evaluate the structure of the metro network: topology, which includes the degree of connection between nodes; geographic coverage; and technical characteristics, which include operational speed and transportation capacity [3]. The goal of this paper is to study the first category, the topological evaluation of metro networks.

Regarding simulating metro passenger flow, random walks are suitable models and many researchers have detected community structures and rankings of nodes in graphs by generating random walks on networks [4]. Given their accessibility, random walks are effective models for network diffusion because they make it simple to analyze how network motions are affected by the network topology [5]. A simple random walk on a network, for example, is a Markov process, which implies that each move only depends on the current position and previous pathways can be ignored. In this situation, transition probabilities can be easily calculated from the numbers and weights of edges between nodes.

Recently, the detection of metro network structure using graph theory and random walks has received a lot of academic attention. The average time it takes to reach a specific metro station for the first time from any other station is negatively correlated with the number of routes that go back to it, according to a study involving 28 real metro systems around the world [6]. The properties of metro networks have been described using graph theory, which allows topological assessments of the routes, structures, and connectivity as well as evaluate the stages of metro development through cluster analysis [7]. Additionally, key nodes in the transportation network can be identified by the connectivity and activity density of the nodes [8]. To predict the inflow and outflow of passengers in metro stations, a new spatio-temporal dynamic graph model has also been proposed. It is also capable of properly capturing both temporal relationship dependencies and dynamic spatial dependencies between different stations [9].

All these studies demonstrate that graph theory and random walks are suitable to be applied to the evaluation of metro network structure. However, there is not yet a thorough investigation of the issue of crowd diffusion from a single point in the metro network. Such problems merit in-depth research moving forward. This paper aims to suggest a method to apply graph theory to the study of such problems of single-point diffusion, which is inspired by a class of generative network models [10]. Simple graphs can be generated stochastically as metro maps through these models, allowing the conclusions to be more general and not specific to any particular existing metro system. Random walks are used to establish the crowd diffusion model, and Monte Carlo method is utilized to determine the general relationship between how crowds diffuse on the metro network and the structure of the network. In different graphs generated by the suggested method, the results show that nodes with degree 2 account for the majority of all nodes, while those with degree 4 or 5 are visited most frequently by random walkers. Furthermore, it emerges that there is a significant linear correlation between the average visiting frequency and degree of nodes.

## 2. Methods

### 2.1. Stochastic generation of simple graphs as metro map

Definitions**:** Denote $G$ as a stochastically generated simple graph. Denote a simple random walk on graph $g$, of length $k$ and starting vertex $v\ as\ SRW(g,k,v)$. A simple random walk visits and moves to a random neighbor repeatedly for iterations equal to its length. $G$ is generated by a modification of **Algorithm 1** [10]. The simulation is only concerned about simple random graphs, so the original Algorithm 1, which concerns multigraphs (connecting already connected vertices and self-edges), will not be used. The modification is also mentioned in the article:

**Algorithm 1** (Random walk model on simple random graphs):

• Fix α ∈ (0, 1]. Also fix a Poisson distribution $P$ on ℕ. The algorithm uses typically the Poisson distribution by $Poisson_{+}(\lambda)$ for $P$ to pick the distance of the random walkers in later steps. The initial graph $G_1$, or when $t = 1$, contains only two connected vertices.

• For $t = 2, \ldots, T$, generate $G_t$ from $G_{t-1}$ from the following steps:

(1) Select uniformly a vertex $V$ from $G_{t-1}$ at random.

(2) With the probability α, attach a new vertex to $V$.

(3) Else, draw a value for length $K \sim P$, a random value from the distribution P, and a vertex $V' \sim SRW(G_{t-1}, V, K)$ with $V'$ the terminal vertex of the walk $SRW(G_{t-1}, V, K)$. Connect $V$ to $V'$ if and only if they are distinct and not connected; else create and attach a new vertex to $V$.

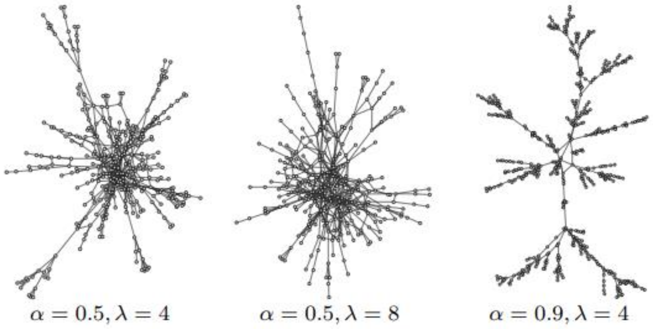
$\alpha = 0.5, \lambda = 4$
$\alpha = 0.5, \lambda = 8$
$\alpha = 0.9, \lambda = 4$

**Figure 1.** Examples of generated simple graphs [10]

Hence, the two variables, α and λ, change the structure of the generated graphs. α is a value between not inclusive 0 and inclusive 1, and λ is the parameter for the Poisson distribution, which denotes the mean number of events within a given interval. The above graphs in Figure 1 [10] depicts the difference of generated graphs with the changing variables α and λ. The difference for changing α is more apparent, and the larger it is, the more sparce the generated graph becomes. Changing λ only affects the number of steps the random walker moves when generating the graph. So, for this part of the simulation, α is set as 0.5, representing the middle value within the range, and λ is set to a randomly chosen number from the normal distribution with average of 6 and a standard deviation of 1, and truncated at 1 and 10 to represent the distribution of the reasonable number of stations traveled by passengers. Generate many simple graphs to be used in the later stages. Each graph represents a randomly generated metro map.

### 2.2. Modeling regular metro passenger flow using random walk

Start with a graph $G$ generated from above, then place uniformly $n_1$ random walkers onto the graph $G$, representing passengers starting their trip from random locations. Run the random walkers $k_1$ steps each, modeling the daily passenger flow of a city metro map. For each vertex $v$ in graph $G$, keep track of the number of visits of that vertex $v$. The numbers for $n_1$ and $k_1$ are chosen from a range. This counts as one simulation.

Repeat this process multiple times on different graphs generated with 2.1 and analyze the topology by examining the relationships between the degree of the vertices and the frequency of visitation.

### 2.3. Diffusion of passengers from one station

Each vertex’s frequency of being visited is kept track of in 2.2, so for the most visited vertex in the same graph $G$ in 2.2, place $n_2$ random walkers on that vertex, representing a situation where many passengers come in at one station. Only one station is used to simplify the simulation, and having many people diffuse from one station is a reasonable depiction of an event in a metro system. Run these random walkers $k_2$ steps. The numbers for $n_2$ and $k_2$ are chosen from a range. This counts as one simulation.

Repeat this process multiple times on the graphs used in 2.2 and analyze the relationship between the number of steps, $k_2$, and the distance from the original vertex.

## 3. Results

By the introduction of method, graphs with different number of nodes are generated. The numbers are: 500, 1000, 2000, 4000, and 8000. Each number group includes 10 generated graphs. Figure 2 illustrates the number of nodes with different degrees. The nodes with a degree of 2 account for the most part of all nodes. As the degree increases, the nodes decrease. The number of nodes with the degree of 1 is almost half the number of 2.

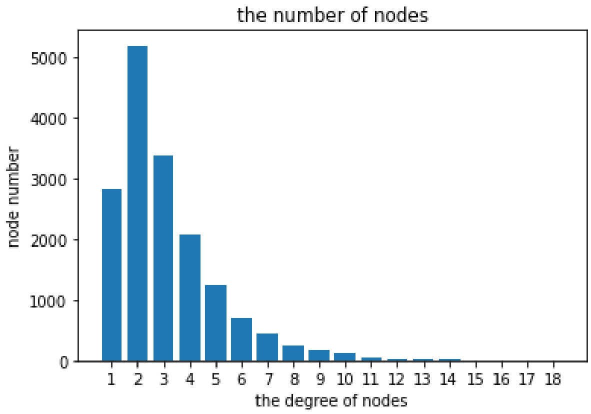

**Figure 2.** Node count for each degree

The distribution is similar to the exponential distribution except for the number 1. The exponential distribution describes the process that the events happen according to a certain probability independently, which is similar to the process that we pick one node and decide whether we add an edge.

Figure 3 demonstrates the frequency of being visited for the nodes with different degrees, depict that the nodes with the degree of 4 or 5 are visited most frequently. Here are some reasons that may matter. The bigger the degree is, the more neighbors the node has, and it is more possible to be visited. At the same time, there are fewer nodes as the degree increase, so there is a balance. In this research, 4 and 5 are the balance point.

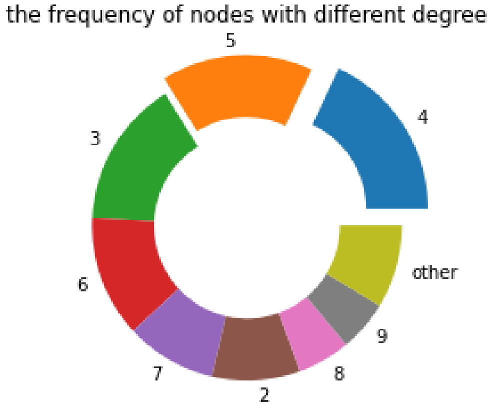

**Figure 3.** Frequency of each degree

Considering the number of the graphs is not uniform and they differ significantly, if the step each random walker will take on different graph are uniform, on those huge graphs, maybe most of the nodes will be not visited for even only once. So, the steps should be adjusted according to the number of the graphs. The steps the walker will take is equal to the number of nodes.

The Figure 4 demonstrates the average frequency of being visited for those nodes with different degree. Figure 4 shows that there is obvious linear relationship between the frequency and the degree. As the degree increases, the frequency increases simultaneously. The linear correlation may result from the method that we use to generate graph. The chance that one will be added a neighbor node is fixed during the whole period, so the connection between nodes is nearly uniformly distributed in the graph to some extent and there is no expectation for nodes with specific degree.

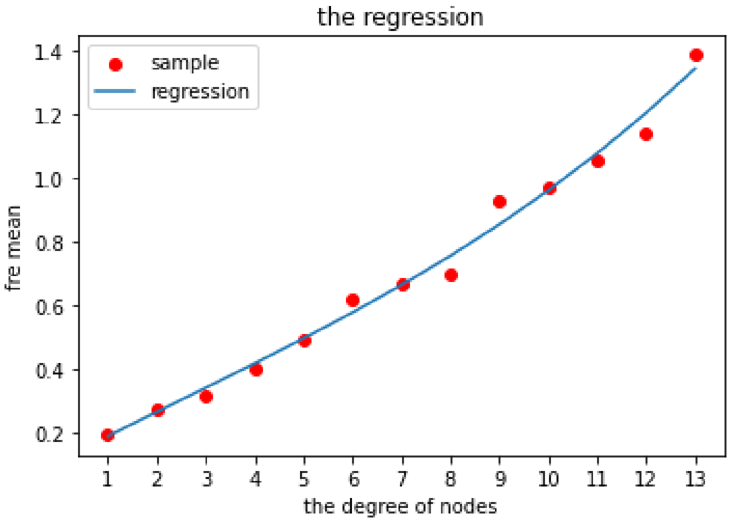

**Figure 4.** Average frequency vs degree

## 4. Conclusion

In this paper, we focused on the random walks on the randomly generated graphs. One method is implemented to generate the graph randomly. Numerical results show that by the method there is few nodes with quite high degree. There is a linear relationship between the average frequency of being visited and the degree of the node. The details regarding the relationship are very likely to related to the value we use in the method.

## References

[1] Chen, H. (2020, December 26). Shanghai adds 7,000th train to Metro fleet. Shanghai Daily. https://www.shine.cn/news/metro/2012252178

[2] Kansky, K. J. (1963). Structure of transportation networks: relationships between network geometry and regional characteristics (Doctoral dissertation, The University of Chicago).

[3] Gattuso, D., & Miriello, E. (2005). Compared analysis of metro networks supported by graph theory. Networks and Spatial Economics, 5(4), 395-414.

[4] Pons, P., & Latapy, M. (2005, October). Computing communities in large networks using random walks. In International symposium on computer and information sciences (pp. 284-293). Springer, Berlin, Heidelberg.

[5] Masuda, N., Porter, M. A., & Lambiotte, R. (2017). Random walks and diffusion on networks. Physics reports, 716, 1-58.

[6] Zhu, Y., Zhao, L., Li, W., Wang, Q. A., & Cai, X. (2016). Random walks on real metro systems. International Journal of Modern Physics C, 27(10), 1650122.

[7] Stoilova, S., & Stoev, V. (2015). An application of the graph theory which examines the metro networks. Transport Problems, 10.

[8] Psaltoglou, A., & Calle, E. (2018). Enhanced connectivity index–A new measure for identifying critical points in urban public transportation networks. International Journal of Critical Infrastructure Protection, 21, 22-32.

[9] Xie, P., Ma, M., Li, T., Ji, S., Du, S., Yu, Z., & Zhang, J. (2022). Spatio-Temporal Dynamic Graph Relation Learning for Urban Metro Flow Prediction. arXiv preprint arXiv:2204.02650.

[10] Bloem-Reddy, B., & Orbanz, P. (2018). Random-walk models of network formation and sequential Monte Carlo methods for graphs. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(5), 871-898.