# Three Stochastic Models On Discrete Structures

**Farkhondeh Alsadat Sajadi**

**Indian Statistical Institute**

**July, 2013**

# Three Stochastic Models On Discrete Structures

**Farkhondeh Alsadat Sajadi**

**Thesis submitted to the Indian Statistical Institute**

**in partial fulfillment of the requirements**

**for the award of the degree of**

**Doctor of Philosophy**

**July, 2013**

**Indian Statistical Institute**

**7, S.J.S. Sansanwal Marg, New Delhi, India.**

*To*

*My Parents and My Sisters*

# Acknowledgements

First and above all, I praise God, the most merciful, for providing me the opportunity to step in one of the most beautiful areas in the world of science, Probability and Statistics. To be able to step strong in this way, I have also been supported by many people to whom I would like to express my sincerest gratitude.

This thesis in its current form, would not have been possible without the guidance, continued support, patience and the help of my thesis advisor, Antar Bandyopadhyay. I would like to sincerely thank him for his warm encouragement, thoughtful guidance, critical comments and correction of the thesis. I am also extremely indebted to him for helping me to accomplish my research work. It has been an honor for me to be his first graduate student.

I wish to express my deepest appreciation and heartfelt thanks to Rahul Roy, a truly great teacher and a wonderful person. Although I call him Rahul-da, as an elder brother, but I am grateful to him for being more like a father to me. I would like to acknowledge him not only for being an excellent teacher, but for his constructive suggestions and criticism and also for offering me valuable advice. His constant support through all stages of my stay in India, excellent care and generous invitations to his house will always remain unforgettable.

Staying In India will always be an unforgettable memory since I had too many wonderful experiences. Describing my visit of India is impossible to sum up in few lines of few pages.

> *"This is indeed India; the land of dreams and romance, of fabulous wealth and fabulous poverty, of splendor and rags, of palaces and hovels, of famine and pestilence, of genii and giants and Aladdin lamps, of tigers and elephants, the cobra and the jungle, the country of a thousand nations and a hundred tongues, of a thousand religions and two million gods, cradle of the human race, birthplace of human speech, mother of history, grandmother of legend, great-grandmother of tradition, whose yesterdays bear date with the mouldering antiquities of the rest of the nations, the one sole country under the sun that is endowed with an imperishable interest for alien prince and alien peasant, for lettered and ignorant, wise*

*and fool, rich and poor, bond and free, the one land that all men desire to see, and having seen once, by even a glimpse, would not give that glimpse for the shows of all the rest of the globe combined".* (Mark Twain, Following the Equator, 1897.)

I arrived in India a little apprehensive about my stay, but all my concerns vanished in due course of time with the assistance of some wonderful people. On the day of my arrival in India, I was warmly received by Tridip Ray, Isha Dewan and Abhay G. Bhatt who remained helpful throughout my stay at ISI. Thanks to all of them.

I would like to express my genuine appreciation to Isha Dewan, the Head of Theoretical Statistics and Mathematics unit of ISI Delhi. I greatly appreciate her excellent assistance, moral support and her helpful advice.

I want to express my sincere gratitude to Rajendra Bhatia and his wife, Irpinder Bhatia. Most of the time, they did not let me to feel homesick. Their moral support and genuine concern of my well being is highly appreciable. I had a wonderful time with them. I learned a great deal about Indian history and culture from them.

I am immensely indebted to the teachers who taught me in ISI, both Delhi and Kolkata. Special thanks to Abhay G. Bhatt, Swagata Nandi, Ravindra B. Bapat, Alok Goswami, Amites Dasgupta and Arijit Chakrabarty. I would also like to thank Anish Sarkar for his courses and also his suggestions related to my research.

I am grateful to K. R. Parthasarathy for useful advice. I would like to acknowledge B.V. Rao. Although I did not have courses with him but I had the chance to interact with him during my stay at ISI Kolkata. He taught me the right techniques and logical procedure to deal with a mathematical problem. I was amazed by his humbleness when he paid me a visit in the hostel when he came to ISI, Delhi.

My sincere appreciation is extended to Arup K. Pal, in-Charge of Students' Academic Affairs, ISI Delhi, for his help and support.

My time in ISI was made enjoyable by many friends who became a part of my life. Though it is impossible to thank everyone individually, I would like to express a few words of gratitude to my best friends. I thank Namrata who greeted me at the door with her warm smiling face on my first day at ISI. With the help of Namrata, Mridu, Amaresh and Rudrani I could adjust to hostel life. I thank all of them. I am also thankful to Rashmi and Vikrant. We had a wonderful time together. I am grateful to Abhimanyu, Debadatta, Ashokankur, Soumendu, Sonal, Anup, Bipul, Ankit, Ankita, Tuhina, Mudra, Kushdesh, Arnab, Kunjesh, Fadikar, Sayan, Upama, Pragya, Susmita, Ridhima, Eshita, Kanika and Anuradha for being such wonderful friends. Special thanks to Priyanka and her family. I am grateful for her help and support during my stay at ISI.

The last three years of ISI life, became more enjoyable with my friend Debleena. I am grateful for the time spent with her and for the excellent discussions with her. Together we have shared so many special moments, contributing to a common story.

Special gratitude to my Iranian friends in India, Mansureh and Sedigheh. Although I have lived far from my family, but communications with them provided the right emotional atmosphere for me. Hereby, I would like to thank them again for everything. I would also like to thank my friend Zohre in Iran for her constant help.

I am thankful to my friends from ISI Kolkata who made my stay there wonderful. Special thanks to Rajat, Neena, Subhajit, Prosenjit, Arunangshu, Tanushree, Momita and Arunabha. I would like to thank my friend Debasish at Dean's office of ISI Kolkata to help me during my stay in India.

I am thankful to all the staff members of ISI Delhi for their help and generosity. Special thanks to V.P. Sharma who was as an administrative officer when I joined ISI. Also thanks to Anil Kumar Shukla, the secretary of Theoretical Statistics and Mathematics unit of ISI Delhi.

I take this opportunity to sincerely acknowledge Indian Statistical Institute for the financial support. I was honored to be an ISI Junior Research fellow for two years and then a Senior Research fellow.

I would like to express gratitude to anonymous reviewers who gave valuable suggestion that has helped to improve the quality of the earlier manuscript.
My sincere thanks are due to the external examiner of my PhD viva, Manjunath Krishnapur for his very valuable remarks on the manuscript which also helped to improve the quality of the earlier manuscript.

Last but by no means least, I would like to pay high regards to my mother and my father, who have given me their unambiguous support throughout as always, for which my mere expression of gratitude does not suffice. I whole-heartedly thank and appreciate them for their love, encouragement, help, tremendous patience and spiritual support in all aspect of my life. Many thanks to them for their faith in me and allowing me to seek my dreams.

Words are not enough to express my deep sense of gratitude towards my lovely sisters, the best ever friends. I want to express my gratitude and deepest appreciation to my elder sister, Fakhri for understanding me so well and helping me so much. I am unable to find words to thank Haniyeh, my younger sister. Leaving home and coming to India to study, would have been impossible without her whole-hearted support and help. She has provided assistance in numerous ways. Finally, I am grateful to my youngest sister, Zeinab for her constant moral support. She was a faraway company for me during my stay away from home.

Farkhondeh Alsadat Sajadi

July 2013

Delhi, India

# Contents

# Chapter 1

# Introduction

In the last three or so decades, the theory of probability has emerged as one of the major tools for studying several natural as well as real world phenomena. In many such cases, the randomness is perhaps assumed artificially but because of the high complexity of these problems, the stochastic modeling has often provided a better understanding than any deterministic model. One such stochastic model which has been of central importance in various applications related to *epidemiology, computer and other electrical networking, combinatorial optimization* and *statistical physics*, is the so called the *random graphs*. Like many other topics of mathematics, the theory of random graphs started with purely mathematical interest, but soon became one of the most important tools to study many applied problems.

In this thesis we will consider the following three problems related to the study of random graphs and stochastic processes defined on them:

(i) Virus spread on a finite network;

(ii) Nearest neighbor algorithm for mean field traveling salesman problem; and

(iii) Random geometric graphs with Cantor distributed vertices.

The first problem is directly related to application of general random graph theory to the spread of a virus or malware in a network which is of interest in epidemiology or computer networking. The second problem is related to a famous combinatorial optimization problem known as *traveling* *salesman problem* and the model we consider arises from statistical physics. We study a specific approximation algorithm for this traveling salesman problem and try to study its performance. The third and the last problem is related to the study of certain types of random graphs. We study a curious case of *random geometric graphs* and show that the standard results may not hold when we differ from the usual assumption in the theory of random geometric graphs. The following section provides more detailed introduction to each of the three problems.

## 1.1 Summary of the thesis

### 1.1.1 Virus spread on a finite network

Our first problem deals with spread of a virus or malware through a network of agents. It involves a very simple *susceptible infected removed (SIR)* model which was studied by Draief, Ganesh and Massouli in (Draief et al., 2008). In this model, each susceptible agent, can be infected by its infected neighbors at a rate, proportional to their number and remains infected till it is removed after an unit time. While it is infected, it has the potential to infect its neighbors. In general, removal can correspond to a quarantine of the machine from the network or patching the machine. In this model, it is assumed that once a node is removed, it is "out of the network". That is, it can no longer be susceptible or infected. Such a model is justified, provided the epidemic spread happens at a rate much faster than the rate of patching of the susceptible machines.

In brief, we consider a virus spread model on a finite closed population of $n$ agents, connected by some neighborhood structure which we model through a graph $G$, where the vertices represent the agents. Starting with some initial infected vertices, at each discrete time step, an infected vertex tries to infect its neighbors with probability $\beta \in (0, 1)$ independently of others and then it dies out. The process continues till all infected vertices die out. Our goal is to find some good approximation to the *total number of infected agents* after the epidemic is over for a general network $G$. To this end, we establish a lower bound for the expected total number of infected agents and show that for a large class of graphs which satisfy certain properties, our lower bound is asymptotically exact. The lower bound is obtained through a graph algorithm, namely, *breadth-first search algorithm* and thus works for any network. We show that the networks for which this approach results to asymptotically exact answer, are the ones which locally "look like a tree". This informal description is made rigorous using the concept of *local weak convergence* described by Aldous and Steele (2004). We also show that our lower bound gives better approximation than the known matrix-based upper bounds which were found by Draief et al. (2008). The details of this problem are available in Chapter 2.

### 1.1.2 Nearest neighbor algorithm for mean field traveling salesman problem

The second problem we study in this thesis, is based on a mix of combinatorial and probabilistic techniques. Graph theory is very much tied to the geometric properties of combinatorial optimization (Avis et al., 2005). The *Traveling Salesman Problem (TSP)*, is an example of a combinatorial optimization problem which has attracted the attention of the mathematicians from ages. The task is to find the *shortest* tour among $n$ cities given the intercity distances. There are several randomized versions of this problem where the distances are taken to be random. In particular the one which attracted considerable attention among mathematicians and computer scientists is known as the *Euclidean TSP*, in which the $n$ cities are randomly distributed in a $d$-dimensional hypercube and the distances between cities are given by the Euclidean metric and are thus random. The other random TSP, which has been of interest within the *statistical physics* community is the *mean field TSP*. Here the distances between pairs of cities, that is, $d(c_i, c_j)$ are taken as independent random variables with a given distribution $F$. Note that in this case, the geometric structure may break since the triangle inequality may not necessarily hold with probability one. In fact we can not quite say that the numbers $d(c_i, c_j)$ really represent distances under any metric. This though seems artificial, but has interest in the statistical physics literature. It is well known in theoretical computer science that given the intercity distances (deterministic or random), the TSP in general is a *NP-Complete* problem (Papadimitriou and Steiglitz, 1998). So there are several approximate algorithms which tries to approximate the optimal tour with polynomial running time. Among them, one of the simplest is the *Nearest Neighbor (NN) Algorithm* (Bellmore and Nemhauser, 1968), which is also known as *the next best method* (GavettBose, 1965). It was one of the first algorithms used to determine an approximate solution to the traveling salesman problem. The algorithm starts with a tour containing a randomly chosen city and then always adds the nearest not yet visited city to the last city in the tour. The algorithm terminates when every city has been added to the tour. For the Euclidean TSP, the performance of this algorithm was studied in (Rosenkrantz et al., 1977), where it has been shown that asymptotically the ratio of the total tour length from NN algorithm to that of the optimal solution is of the order $\log n$. For the mean field set up in a recent work of Wästlund (2010), it is shown that if the underlying distribution of the intercity vertices has a density near the origin which has a non-zero limit at $0$, then the total length of the optimal tour is asymptotically constant. In Chapter 3, we show that under same assumption the total length of NN tour is asymptotically almost surely of the order $\log n$. This shows that the performance of the NN algorithm in comparison to the optimal is same in both mean filed and Euclidean set ups. Moreover we also consider general distribution function for the i.i.d. intercity distances and show that the asymptotic behavior of the total length of NN tour depends on the limiting properties of the density function near $0$.

### 1.1.3 Random geometric graphs with Cantor distributed vertices

The third and the last problem of the thesis considers a special *random geometric graph*. The theory of random graphs was established in the late fifties and early sixties of the last century. Among a few papers which appeared around (and even before) that time, the paper by Erdős and Rényi (1960) is generally considered to have founded the field of random graphs. The authors Erdös and Rényi studied the following random graph, which is now named after them: it is a graph with $n$ vetrices where an edge is present with probability $p$ independent of other edges. Another known model of random graphs is *random geometric graph (RGG)*. This graph is obtained by placing $n$ vertices independently according to a common distribution on Euclidean space and connecting two vertices if and only if, they are within some specified critical distance. One of the main aspect what one studies in here is the connectivity of this graph. For uniform and non-uniform underlying distribution, there are results on the connectivity threshold. Appel and Russo (1997) proved strong law results for graphs, constructed on independent random variables distributed uniformly on $[0,1]^d$. Penrose (1999) extended this to graphs where vertices are independent random points in $\mathbb{R}^d, d \geq 2$ with common density having connected compact support with smooth boundary. He also assumed that, the essential infimum of density over this support, is positive. Sarkar and Saurabh (2010) studied the weak convergence of connectivity threshold when the density $f$ of the underlying distribution on $[0,1]$, is regularly varying at the origin. In Chapter 4, we give asymptotic result for connectivity of random geometric graphs, where the underlying distribution of the vertices has no density. For that, we consider $n$ independent and identically *Cantor* distributed points on $[0,1]$. We show that for this random geometric graph, the connectivity threshold $R_n$, converges almost surely to a constant $1-2\phi$ where $0 < \phi < 1/2$, which for the standard Cantor distribution is $1/3$. We also show that $\|R_n - (1-2\phi)\|_1 \sim 2\, C\,(\phi)\; n^{-1/d_\phi}$ where $C\,(\phi) > 0$ is a constant and $d_\phi := -\log 2/\log \phi$ is the *Hausdorff dimension* of the generalized Cantor set with parameter $\phi$.

In the next section, we present some graph theoretical concepts which we use in the chapters that follow. Each chapter, is devoted to one of three problems that we mentioned above. In Chapter 2, we study the first problem which is about the spreading of a virus on finite networks. The details of this problem are based on (Bandyopadhyay and Sajadi, 2012a). The second problems, based on (Bandyopadhyay and Sajadi, 2013), is described in Chapter 3 and involves the application of NN algorithm for the mean field TSP. Third problem which is on RGG with Cantor distributed vertices, is discussed in Chapter 4 and it is based on (Bandyopadhyay and Sajadi, 2012b).

## 1.2 Preliminaries

### 1.2.1 Graph-theoretical terminology

The theory of graph began in 1735, when Leonhard Euler solved a popular puzzle about Königsberg's bridges (Alexanderson, 2006). The city of Königsberg (now is known as Kaliningrad) included two large islands and there were seven bridges that join different parts of this city. The puzzle was to find a way to walk through the city that wouldn't cross each bridge twice. The field of graph theory has exploded after Euler solved this problem and became a very popular area of discrete mathematics. Graph theory can be partitioned into two parts: the areas of undirected graphs and directed graphs (digraphs). Even though both areas have important applications, for various reasons, undirected graphs have been studied much more extensively than directed graphs (Bang-Jensen and Gutin, 2009). In this thesis, we shall focus on undirected graphs. In the following, we provide most of the terminology and notation used in this thesis.

An *undirected graph* (or just graph) $G$ consists of a non-empty countable set $V(G)$ of elements called vertices and a countable set $E(G) \subseteq V \times V$ called edges. Each edge $e = \{u, v\} \in E(G)$ is an unordered pair of distinct vertices $u$ and $v$, which are declared to be adjacent or neighbors. We write $G = (V, E)$ which means that $V$ and $E$ are the vertex set and edge set of $G$, respectively.

A *directed graph* (or digraph) is a graph whose edges have direction and are called arcs. Arrows on the arcs are used to encode the directional information. Thus, an arc from vertex $u$ to vertex $v$ indicates that one may move from $u$ to $v$ but not from $v$ to $u$.

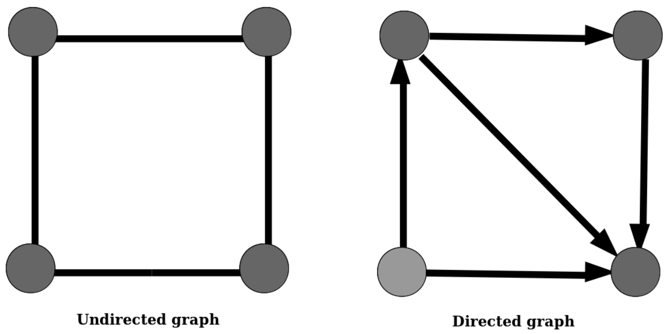


Figure 1.1: A graph and a digraph

A *subgraph* $G_0$ of a graph $G$, is a graph whose vertex set $V_0$, is a nonempty subset of the vertices of $G$ and whose edges are a subset of the edges of $G$.

The cardinality of the set of neighbors of $u$ is called the *degree* of $u$. When the degree of every vertex is finite, we say that $G$ is *locally finite*. When the set $V$ itself is finite, we say that $G$ is finite. A *path* is a sequence of consecutive edges in a graph and the length of the path is the number of edges traversed. Two vertices in a graph are said to be *connected* if there is a path that begins at one and ends at the other. The *graph distance* from $u$ to $v$ is then defined as the minimum length of a path from $u$ to $v$. *Being connected to* is an equivalence relation on $V$; the associated equivalence classes are called the *connected components* of $G$. When there is only one connected component, we say that $G$ is connected.

Graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ are *isomorphic*, denoted $G_1(V_1, E_1) \cong G_2(V_2, E_2)$, if there is a bijection (one-to-one correspondence) $\psi$ from $V_1$ to $V_2$ such that any two vertices $u$ and $v$ of $G_1$ are adjacent in $G_1$ if and only if $\psi(u)$ and $\psi(v)$ are adjacent in $G_2$.

### 1.2.2 Graph algorithms

An algorithm is any well-defined computational procedure that takes a set of values, as input and produces a set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output (Cormen et al., 2009). In the following, we briefly describe two different graph algorithms.

**Breadth-first search** *Breadth-first search* (BFS) is one of the simplest algorithms for searching a graph and the archetype for many important graph algorithms (Cormen et al., 2009). Consider a graph $G = (E, V)$ and a distinguished *root* vertex $v_0 \in V$. A BFS with the start point $v_0$ is as follows. First it explores all vertices which are adjacent to $v_0$. In fact, it discovers every vertex which is at graph distance one from $v_0$, namely $\{v_1, v_2, \ldots, v_l\}$. Then for each $i = 1, 2, \ldots, l$ it explores all unvisited neighbors of $v_i$. These new visited vertices are at graph distance 2 from $v_0$. The search continues in this fashion until it reaches all vertices which are reachable from the root $v_0$. The name of BFS for this algorithm is because, all vertices at distance $k$ from $v_0$ are discovered before discovering any vertices at distance $k + 1$. BFS traverse a connected component of a given graph and makes a spanning tree out of that graph with root $v_0$ (see Figure 1.2 for an example). In BFS spanning tree, for any vertex $u$ reachable from $v_0$, the simple path from $v_0$ to $u$, corresponds to a "shortest path" from $v_0$ to $u$, that is, a path containing the smallest number of edges. BFS algorithm is used for both directed and undirected graphs. We briefly describe the algorithm here.

`Step-0 Input graph` $G$ `with a linear ordering of its vertices, say`

$V := \{v_0, v_1, v_2, \cdots, v_{n-1}\}$. Let $T \leftarrow \{v_0\}$ and $N \leftarrow \{v_0\}$.

Step-1 Write $N = \{v_{i_1}, v_{i_2}, \cdots, v_{i_r}\}$ for some $r \geq 1$ such that

$$i_1 < i_2 < \cdots < i_r \,.$$

Step-2 For $l = 1$ to $r$ find all neighbors $u$ of $v_{i_l}$ which are not in $T$, put $N' \leftarrow \left\{u \,\middle|\, u \sim v_{i_l} \text{ and } u \notin T\right\}$ and update $T$ as

$$T \leftarrow T \cup N'.$$

Step-3 Update $N \leftarrow N'$.

Step-4 Go to Step-1 unless vertex set of $T$ is same as that of $V$.

Step-5 Stop with output $T$ as the BFS spanning tree with root $v_0$.

Note that the BFS spanning tree is not necessarily unique, it depends on the starting point $v_0$ which is typically called the root and also it depends on the ordering of the vertices in which the exploration of neighbors is done in Step-2. Also note that if $G$ is a tree to start with, then BFS spanning tree is just itself. Figure 1.2 provides an illustration.

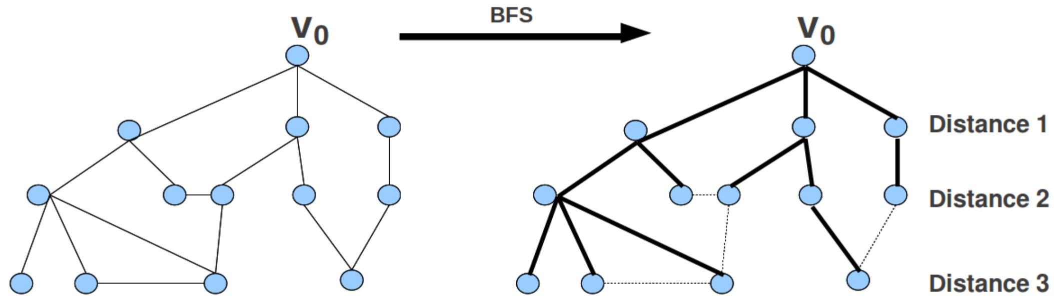


Figure 1.2: BFS Algorithm

In chapter 2, we show an application of BFS algorithm to get a lower bound on the expected number of ever infected vertices.

**The Nearest Neighbor algorithm** In mathematics and computer science, an optimization problem refers to an attempt to minimize or maximize a real function so called, the objective function. For example consider TSP in which a salesman visits $n$ cities cyclically. He visits each city only once, and finishes up where he started. In this case, the typical question which arises, in what order he should visit the cities to minimize the distance traveled. Although there are optimal algorithms to answer this question, but it is computationally unfeasible to obtain the optimal solution to TSP. In fact, if number of cities is large, then it is almost impossible to have an optimal solution within a reasonable amount of time. Therefore to solve such problems, instead of optimal algorithms, one can use heuristics ones. Here we mention to two shortest path algorithms, namely *greedy*(GR) and *nearest neighbor* (NN) algorithms as heuristic algorithms which they are used to get a solution near to optimal one. It is known that, for TSP on $n$ cities, the running time for NN algorithm is $O(n^2)$. The implementation time of the GR algorithm is $O(n^2 \log_2 n)$ and is thus somewhat slower than NN (Johnson and McGeoch, 1997). Every decision which the GR algorithm takes, is the one with the most obvious immediate advantage. For the TSP on $n$ cities, which are labeled as $\{c_1, c_2, \dots, c_n\}$, this algorithm works as follows. First it sorts all the edges $\{c_i, c_j\}$. Then repeatedly, it selects the shortest edge and adds it to the tour as long as it doesn't create a cycle with less than $n$ edges. The other heuristic algorithm is NN algorithm which is one of the first algorithms used to determine an approximate solution to the TSP. For each edge $\{c_i, c_j\}$, let $d(c_i, c_j)$ be the distance between city $c_i$ and city $c_j$. Briefly, in the NN algorithm, a tour is constructed as follows:

Step-0: Input graph $G$ with a linear ordering of its vertices say $V := \{c_1, c_2, \dots, c_n\}$. Let $Tour \leftarrow \{c_1\}$ and $c_{\pi(1)} = c_1$.

Step-1: Write $Tour \leftarrow \left\{c_{\pi(1)}, c_{\pi(2)}, \dots, c_{\pi(i)}\right\}$. Choose $c_{\pi(i+1)}$ to be the city $c_j$ that minimizes

$$\{d(c_{\pi(i)}, c_j) : j \neq \pi(k), 1 \leq k \leq i\}.$$

Update $Tour$ as $Tour \leftarrow Tour \cup \left\{c_{\pi(i+1)}\right\}$.

Step-2: Go to Step-1 unless $V \setminus Tour = \emptyset$.

Step-3: Stop with output $Tour$ as the NN tour with starting city $c_1$.

For the convenience, when there are ties in `Step-1`, we assume that they can be broken arbitrarily. The NN algorithm can be improved by repeating the algorithm for each possible starting city and then take the minimum solution among them (GavettBose, 1965). Figure 1.3 shows an example of using NN algorithm for finding the shortest tour among 5 cities. Starting city is $c_1$ and the next visited cities in order are : $c_3, c_5, c_4$ and $c_2$.

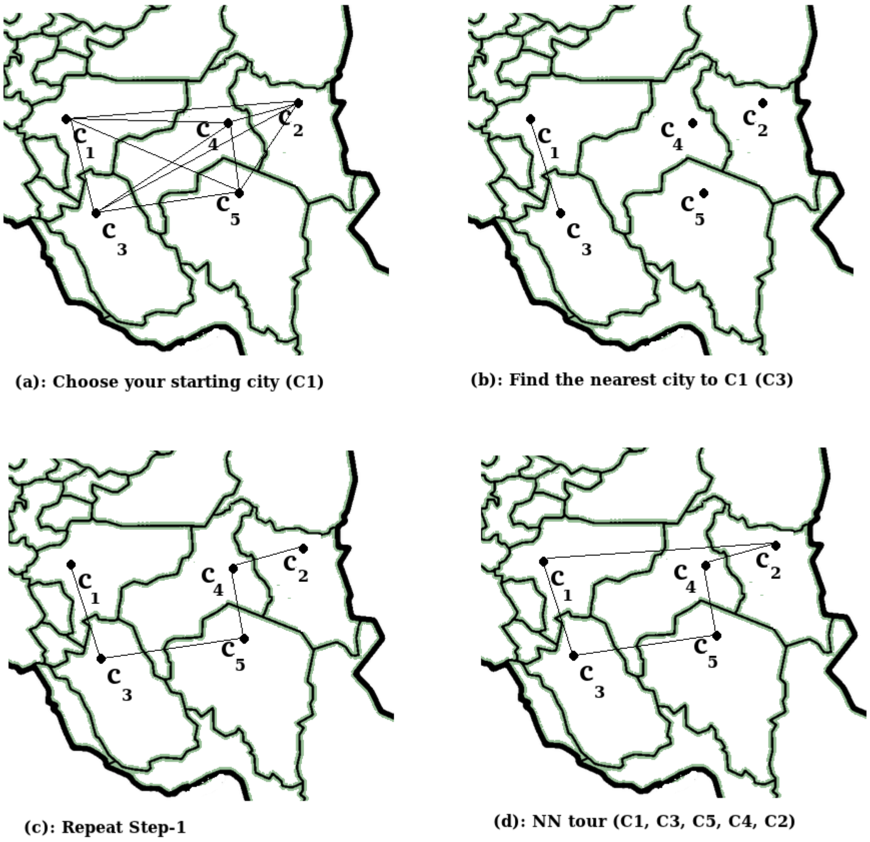


Figure 1.3: The nearest neighbor tour

In Chapter 3 we present an application of NN algorithm for the mean filed TSP.

### 1.2.3 Connectivity threshold of random graphs

As we mentioned earlier, the theory of random graphs began in the late 1950s in several papers by Erdös and Rényi. Random graphs are often used as a model of real-world networks such as social links, computer networks, the Internet, the biological networks and the linking structure of the World Wide Web (Barabási et al., 2003, Gilbert, 1961, Newman et al., 2002, Penrose, 2003, Watts and Strogatz, 1998).

Let $\| \, . \, \|$ be some norm on $\mathbb{R}^d$, for example the Euclidean norm and let $r$ be some positive parameter. A *geometric graph* on a finite set $V \subset \mathbb{R}^d$, is an undirected graph with vertex set $V$ and with undirected edges, connecting all those pairs $\{u, v\}$ such that $\| \, u - v \, \| \, \leq r$. Other terms which have been used for geometric graph are interval graphs (when $d$ = 1), disk graphs (when $d$ = 2), and proximity graphs (Penrose, 2003). *Random geometric graph* (RGG), is a geometric graph on random point configurations (Gilbert, 1961, Penrose, 1997, 2003). Often the vertices of RGG are assumed to be distributed on $[0, 1]^d$ according to a Poisson point process. Denote RGG by $\mathcal{G} = \mathcal{G}(V_n, r)$. The connectivity threshold $R_n$ for a finite set $V_n \subset \mathbb{R}^d$, defined to be the minimum value of $r$ such that $\mathcal{G}$ is connected. $R_n$ for $V_n$ is also, the longest edge length of the minimal spanning tree on $V$; see for example (Penrose, 1997). It has been shown that if $r \geq \sqrt{\frac{\log n + \gamma_n}{\pi n}}$ then $\mathcal{G}$ is connected with high probability as $n \to \infty$ if and only if $\gamma_n \to +\infty$ and disconnected with high probability if and only if $\gamma_n \to -\infty$ (Gupta and Kumar, 1998, Penrose, 1997). We study the *connectivity threshold* of one particular RGG in Chapter 4.

### 1.2.4 Cantor distribution

*The Cantor set* $C$, is a rather remarkable subset of $[0, 1]$, which was first discovered by Smith (1875) but became popular after Cantor (1883). There are different ways to define and construct the Cantor set. But, the popular one is the Cantor middle-thirds or ternary set construction. The resulting set, is called the *Standard Cantor* set, which is constructed on the interval $[0, 1]$ as follows. One successively removes the open middle third of each subinterval of the previous set. The Cantor set itself is the infinite intersection of all remaining sets. More precisely, starting with $C_0 := [0, 1]$, we inductively define

$$C_{n+1} := \bigcup_{k=1}^{2^n} \left( \left[ a_{n,k}, a_{n,k} + \frac{b_{n,k} - a_{n,k}}{3} \right] \cup \left[ b_{n,k} - \frac{b_{n,k} - a_{n,k}}{3}, b_{n,k} \right] \right)$$

where $C_n := \bigcup_{k=1}^{2^n} [a_{n,k}, b_{n,k}]$. The Standard Cantor set is then defined as

$$C = \bigcap_{n=o}^{\infty} C_n$$

Figure 1.4 shows the Cantor ternary set which is created by repeatedly deleting the open middle thirds of a set of line segments.

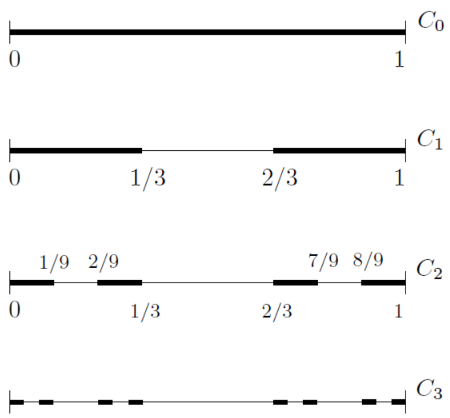


Figure 1.4: Construction of the standard Cantor set

For constructing the one-dimensional generalization of the Cantor set, we start with unit interval $[0, 1]$ and at first stage we delete the interval $(\phi, 1 - \phi)$ where $0 < \phi < 1/2$ . Then, this procedure is reiterated with two segments $[0, \phi]$ and $[1 - \phi, 1]$. We continue ad infinitum.

Lad and Taylor (1992) have defined a probability distribution based on the Cantor set. The *Cantor distribution* with parameter $\phi$ where $0 < \phi < 1/2$ is the distribution of a random variable $X$ defined by

$$X = \sum_{i=1}^{\infty} \phi^{i-1} Z_i \tag{1.2.1}$$

where $Z_i$ are i.i.d. with $\mathbb{P}[Z_i = 0] = \mathbb{P}[Z_i = 1 - \phi] = 1/2$. Intuitively one can construct this distribution on the interval $[0, 1]$, as follows. Start with unit probability mass uniformly distributed over $[0, 1]$. After deleting the interval $(\phi, 1 - \phi)$, by rescaling, make the total probability mass to be one. Continue this procedure to infinity. Note that at $n^{\text{th}}$ stage the probability mass is uniformly distributed over $2^n$ compact intervals each of length $\phi^n$. If a random variable $X$ admits a representation of the form (1.2.1), we will write $X \sim Cantor(\phi)$, and will say that $X$ has a Cantor distribution with parameter $\phi$. Note that for $\phi = 1/3$ we obtain the *standard Cantor distribution*, in which unit probability mass is concentrated on those points in $[0,1]$ whose ternary expansion contains only the digits 0 and 2. In the other words, the standard Cantor distribution is the distribution that is uniform on the standard Cantor set. Observe that Cantor $(\phi)$ is self-similar, in the sense that,

$$X \stackrel{d}{=} \begin{cases} \phi X & \text{with probability } 1/2 \\ \phi X + 1 - \phi & \text{with probability } 1/2 \end{cases} \tag{1.2.2}$$

This follows easily by conditioning on $Z_1$.

# Chapter 2

# Virus spread on finite networks

## 2.1 Introduction

### 2.1.1 Background and Motivation

Often it is observed that the normal operation of a system which is organized in a network of individual machines or agents, is threatened by the propagation of a harmful entity through the network. Such harmful entities are often termed as *viruses*. For example the Internet, as a network is threatened by the computer viruses and worms which are self-replicating pieces of code, that propagate in a network of computers. These codes use a number of different methods to propagate, for example an e-mail virus typically sends copies of itself to all addresses in the address book of the infected machine. Weaver et al. (2003) gives a good survey of different techniques of propagation for computer viruses.

The progress of virus spread, through the network is amenable to mathematical modeling. Such models, can be used to explain patterns or predict the future outcome of an epidemic process. The study of mathematical models for epidemic spread has a long history in biological epidemiology and in the study of computer viruses. Although the first model for epidemic spread, is more than a century old (Hamer, 1906), one of the simplest and most fundamental of all epidemiological models, is the one due to work of Kermack and McKendrick (1927), where they introduced the first stochastic theory for epidemic spread. They proved the existence of an epidemic threshold, which determines whether the epidemic will spread or die out. They introduced the so-called "SIR model", in which individuals can be classified by their epidemiological status, *susceptible infected removed (SIR)*. In this model, every vertex is either infected or healthy (but susceptible). Each susceptible agent, can be infected by its infected neighbors at a rate, proportional to their number and remains infected till it is removed after an unit time. While it is infected, it has the potential to infect its neighbors. In general, removal can correspond to a quarantine of the machine from the network or patching the machine. In this model, it is assumed that once a node is removed, it is "out of the network". That is, it can no longer be susceptible or infected. Such a model is justified, provided the epidemic spread happens at a rate much faster than the rate of patching of the susceptible machines. As mentioned in Draief et al. (2008), earlier work mainly focused on finding or approximating the *law of large numbers* limit where the stochastic behavior was approximated by its mean behavior and hence mainly studied deterministic models. More recent works (Barbour and Utev, 2004, Lefèvre and Utev, 1995), have focused on stochastic nature of the models and have tried to prove asymptotic distribution of the number of survivors, using a key concept called *basic reproductive number*, which is defined as the expected number of secondary infective, caused by a single primary infective. This concept of basic reproductive number is well defined under the *uniform mixing* assumption, that is, when any infective can infect any susceptible equally likely and hence the underlying network is given by a complete graph. For a general network, where basic reproductive number may become vertex dependent, it is not clear how to use this concept effectively. In this chapter, we study this model on a general network.

### 2.1.2 Model

We consider a closed population of $n$ agents, connected by a network structure, given by an undirected graph $G = (V, E)$ with vertex set $V$ containing all the agents and edge set $E$. A vertex can be in either of the three states, namely, *susceptible (S)*, *infected (I)* or *removed (R)*. At the beginning, the initial set of infected vertices is assumed to be non-empty and all others are susceptible. The evolution of the epidemic is described by the following discrete time model:

- After a unit epoch of time, each infected vertex instantaneously tries to infect each susceptible neighbor with probability $\beta \in (0, 1)$ independent of all others.

- Each infected vertex is removed from the network after a unit time.

Mathematically, at an integer multiple of unit time, say $t$, if a susceptible vertex $v$ has $I_v(t)$ neighbors who are infected, then the probability of $v$ being infected instantaneously will be $1 - (1 - \beta)^{I_v(t)}$ and each susceptible vertex will get infected independently. Also an infected vertex remains in the network only for a unit time, after that it tries to infect its susceptible neighbors and then it is immediately removed.

As pointed out by Draief et al. (2008), this is a simple model, falling in the class of models known as Reed-Frost Models, where infection period is deterministic and is same for every vertex. It is worth noting that the evolution of the epidemic can be modeled as a Markov chain.

It is interesting to note that, the model is essentially same as the i.i.d. Bernoulli bond percolation model with parameter $\beta$ (Grimmett, 1999). This is because the set of ever infected (or removed) vertices is same as the union of connected open components of i.i.d. bond percolation on $G$, containing all the initial infected vertices. Although for percolation, it is customary to work with an infinite graph $G$. If $G$ is the complete graph $K_n$, then this model is fairly well studied in literature and is known as the *binomial random graph*, also known as Erdös-Rényi random graph (Bollobás, 2001, Janson et al., 2000).

In this chapter, our goal is to study the total number of vertices that eventually become infected (and hence removed) without specifying the underlying network. In the paper by Draief et al. (2008), the authors derived an explicit upper bound of the expected number of vertices ever infected which depends on both the size of the network as well as the infection rate $\beta$. These bounds also needed an assumption of "small" value for $\beta$. Unfortunately, the work of Draief et al. (2008) did not provide any indication whether the derived upper bound is a good approximation of the quantity of interest. In this work, we derive a simple lower bound of the expected number of vertices ever infected which works for every infection rate $0 < \beta < 1$. Our lower bound is based on the *breadth-first search (BFS)* algorithm and hence easily computable for any general finite network $G$. We also prove that, under certain assumptions on the qualitative behavior of the underlying graph, namely if $G$ "*locally looks like a tree*" in the sense of Aldous and Steele (2004) *local weak convergence*, then our lower bound is asymptotically exact for "small" $\beta$, thus it provides a good approximation when the network is "large". As we will see later, for such graphs $G$, the range we cover for $\beta$ always includes the range in which the upper bound obtained by Draief et al. (2008) holds and in all these cases, the upper bound over estimates the expected total number of infections.

### 2.1.3 Outline

In the following section, we state and prove our main results. Section 2.3 gives several examples where our lower bound holds and gives asymptotically correct answer. Finally in Section 2.4 we summarize the merits of our work and indicate some of its limitations as well.

## 2.2 Main Results and Proofs

We will denote by $Y^{G,I}$, the total number of vertices ever infected when the epidemic runs on a network $G$ and the infection starts at the vertices in $I \subseteq V$. Note that $Y^{G,I}$ implicitly depends on the size of the network. In Subsection 2.2.1 we present the results, when the epidemic starts with only one infected vertex. We generalize these results for epidemic starting with more than one infection, which are presented in Subsection 2.2.2. In both cases, our results rely on *breadth-first search (BFS)* algorithm, which has been described in Subsection 1.2.2. Before stating our main results, since we will compare our lower bound of $\mathbb{E}[Y^{G,I}]$ with the upper bound obtained in Draief et al. (2008), we present here two main theorems from their work. Let $A$ denote the adjacency matrix of the undirected graph $G$ and $\lambda_1(A)$, the eigenvalue with the largest absolute value.

**Theorem 2.2.1** (Draief et al., 2008, Theorem 2.1)**.** *Suppose* $\beta\lambda_1(A) < 1$. *Then,*

$$\mathbb{E}[Y^{G,I}] \leq \frac{1}{1-\beta\lambda_1(A)}\sqrt{n|I|} \tag{2.2.1}$$

*where $I$ is the set of vertices initially infected.*

**Theorem 2.2.2** (Draief et al., 2008, Theorem 2.3)**.** *Let $G$ be an arbitrary graph with maximal node degree denoted by $\Delta$. If $\beta\Delta < 1$ then*

$$\mathbb{E}[Y^{G,I}] \leq \frac{1}{1-\beta\Delta}|I|\,. \tag{2.2.2}$$

### 2.2.1 Starting with only one infected vertex

Our first result gives a lower bound of the expected total number of vertices ever infected, starting with exactly one infected vertex.

**Theorem 2.2.3.** *Let $G$ be an arbitrary finite graph and $v_0 \in V$ be a fixed vertex of it. Let $T$ be a spanning tree of the connected component of $G$ containing the vertex $v_0$ and rooted at $v_0$. Let $Y^{T,\{v_0\}}$ be the total number of vertices ever infected when the epidemic runs only on $T$ and starting with exactly one infection at $v_0$. Then*

$$\mathbb{E}\left[Y^{T,\{v_0\}}\right] \leq \mathbb{E}\left[Y^{G,\{v_0\}}\right] \text{ for all } 0 < \beta < 1\,. \tag{2.2.3}$$

*Moreover, if $\mathcal{T}$ is a BFS spanning tree of the connected component of $v_0$ rooted at $v_0$, then*

$$\mathbb{E}\left[Y^{T,\{v_0\}}\right] \leq \mathbb{E}\left[Y^{\mathcal{T},\{v_0\}}\right] \leq \mathbb{E}\left[Y^{G,\{v_0\}}\right] \text{ for all } 0 < \beta < 1\,. \tag{2.2.4}$$

*Proof.* Suppose $G = (V, E)$ where $V$ is the set of vertices and $E$ is the set of edges and let $H = (V, E')$ where $E' \subseteq E$. So $H \subseteq G$, is a spanning sub-graph of $G$. Note that $v_0$ is a vertex in both $H$ and $G$. Let $(X_e)_{e\in E}$ be i.i.d. Bernoulli $(\beta)$ random variables indexed by the edges of the graph $G$. We consider the random graphs $G_\beta := (V_\beta, E_\beta)$ and $H_\beta := \left(V_\beta, E'_\beta\right)$ with the same vertex set $V_\beta = V$ and the random sets of edges $E_\beta := \left\{e \in E \,\middle|\, X_e = 1\right\}$ and $E'_\beta := \left\{e \in E' \,\middle|\, X_e = 1\right\}$. Note that $H_\beta$ is a spanning sub-graph of $G_\beta$. Let $C^{G,v_0}$ and $C^{H,v_0}$ be the connected components of the vertex $v_0$ in $G_\beta$ and $H_\beta$ respectively. From definition $C^{H,v_0} \subseteq C^{G,v_0}$.

Now it follows from the definition of the infection spread model that $\left|C^{G,v_0}\right| \stackrel{d}{=} Y^{G,\{v_0\}}$

and $\left|C^{H,v_0}\right| \stackrel{d}{=} Y^{H,\{v_0\}}$. So to prove equation (2.2.3) observe that

$$\mathbb{E}\left[Y^{T,\{v_0\}}\right] = \mathbb{E}\left[\left|C^{T,\{v_0\}}\right|\right] \leq \mathbb{E}\left[\left|C^{G,\{v_0\}}\right|\right] = \mathbb{E}\left[Y^{G,\{v_0\}}\right] .$$

For the second part, we note that if $T$ is a spanning tree of $G$ with root $v_0$, then $d_G\left(v,v_0\right) \leq d_T\left(v,v_0\right)$ for all $v \in V$, where $d_G$ and $d_T$ are the graph distance functions on $G$ and $T$ respectively. Moreover, the BFS algorithm preserves the distances, so if $\mathcal{T}$ is a BFS spanning tree with root $\{v_0\}$ then we must have

$$d_G\left(v,v_0\right) = d_{\mathcal{T}}\left(v,v_0\right)$$

for all $v \in V$. Thus $d_{\mathcal{T}}\left(v,v_0\right) \leq d_T\left(v,v_0\right)$ for all $v \in V$. Now from the model description, it follows that for any spanning tree $T$ with root $v_0$ we have

$$\mathbb{E}\left[Y^{T,\{v_0\}}\right] = \sum_{v\in V} \beta^{d_T(v,v_0)} .$$

So we conclude that

$$\mathbb{E}\left[Y^{T,\{v_0\}}\right] = \sum_{v\in V} \beta^{d_T(v,v_0)} \leq \sum_{v\in V} \beta^{d_{\mathcal{T}}(v,v_0)} = \mathbb{E}\left[Y^{\mathcal{T},\{v_0\}}\right] ,$$

as $0 < \beta < 1$. □

Let $\mathrm{LB}^{G,\{v_0\}} := \mathbb{E}\left[Y^{\mathcal{T},\{v_0\}}\right]$ be the lower bound obtained through BFS algorithm for a BFS spanning tree $\mathcal{T}$ of $G$, rooted at $v_0$. Then from the proof of Theorem 2.2.3 we get that

$$\mathrm{LB}^{G,\{v_0\}} = \sum_{v\in V} \beta^{d_G(v,v_0)} , \tag{2.2.5}$$

which is free of the choice of the BFS spanning tree. Later, we will see that, this helps us to generalize the lower bound for epidemic starting with more than one infected vertex. We also note that $\mathrm{LB}^{G,\{v_0\}}$ can be easily computed using the breadth-first search algorithm described earlier.

Our next result shows that if we have a "large" finite graph $G$ on $n$ vertices and the epidemic starts with exactly one infected vertex $v_0$, such that any cycle containing $v_0$ is "relatively large", that is of order $\Omega(\log n)$, then the lower bound $\mathrm{LB}^{G,\{v_0\}}$ given above, is asymptotically same as the exact quantity $\mathbb{E}\left[Y^{G,\{v_0\}}\right]$.

To state the result rigorously, we use the following graph theoretic notations. Given a graph $G$, a fixed vertex $v_0$ of $G$ and $d \geq 1$, let $V_d(G)$ be the set of vertices of $G$ which are at a *graph distance* at most $d$ from $v_0$ in $G$. Let $N_d(G, v_0)$ be the induced sub-graph of $G$ on the vertices $V_d(G)$.

**Theorem 2.2.4.** *Let $G_n$ be a connected graph on $n$ vertices and $\{(G_n, v_0^n)\}_{n\geq 1}$ be a sequence of rooted graphs with roots $\{v_0^n\}_{n\geq 1}$ such that there exists a sequence $\alpha_n = \Omega(\log n)$ with $N_{\alpha_n}(G_n, v_0^n)$ is a tree for all $n \geq 1$. Then, there exists $0 < \beta_0 \leq 1$, such that for all $0 < \beta < \beta_0$*

$$\left|\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] - LB^{G_n,\{v_0^n\}}\right| \longrightarrow 0 \;\; as \;\; n \to \infty \tag{2.2.6}$$

*and therefore* $\dfrac{\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right]}{LB^{G_n,\{v_0^n\}}} \longrightarrow 1$ *as* $n \to \infty$.

*Proof.* Let $\mathcal{T}_n$ be a BFS spanning tree rooted at $v_0^n$ of the graph $G_n$ and as defined earlier let $\mathrm{LB}^{G_n,\{v_0^n\}} = \mathbb{E}\left[Y^{\mathcal{T}_n,\{v_0^n\}}\right]$. Denote $\partial^*_{\alpha_n} N_{\alpha_n}(G_n, v_0^n)$ the set of infected vertices in $G_n$ after $\alpha_n$ units of time starting with one infected vertex $v_0^n$. Then

$$\begin{aligned}
\mathrm{LB}^{G_n,\{v_0^n\}} &\leq \mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] \\
&\leq \mathbb{E}\left[Y^{N_{\alpha_n}(G_n,v_0^n),\{v_0^n\}}\right] + n\mathbb{E}\left[\left|\partial^*_{\alpha_n} N_{\alpha_n}(G_n, v_0^n)\right|\right] \\
&\leq \mathbb{E}\left[Y^{N_{\alpha_n}(G_n,v_0^n),\{v_0^n\}}\right] + n^2\beta^{\alpha_n} \\
&\leq \mathrm{LB}^{G_n,\{v_0^n\}} + n^2\beta^{\alpha_n}.
\end{aligned} \tag{2.2.7}$$

Note that the first term of the second inequality in (2.2.7) is the expected number of infected nodes within an $\alpha_n$ neighbourhood of the initial infective $v_0^n$. The second term there is an upper bound of the expected number of nodes which may become infected outside of $\alpha_n$ neighbourhood of $v_0^n$. But the number of nodes outside the neighbourhood is bounded by

$n - \mathbb{E}\left[Y^{N_{\alpha_n}(G_n, v_0^n), \{v_0^n\}}\right] \leq n$. For the third equality note that since we have assumed that $N_{\alpha_n}(G_n, v_0^n)$ is a tree, so the nodes which are on the boundary of $\alpha_n$ neighbourhood of $v_0^n$, that is the infected vertices in $G_n$ after $\alpha_n$ units of time starting with one infected at vertex $v_0^n$, have probability $\beta^{\alpha_n}$ to get infected after $\alpha_n$ units of time. The last inequality follows from the fact that $N_{\alpha_n}(G_n, v_0^n)$ is a tree and hence is a subtree of $\mathcal{T}_n$. This proves (2.2.6) since by assumption $\alpha_n = \Omega(\log n)$. The last part of the theorem follows from the fact that $\mathrm{LB}^{G_n, \{v_0^n\}} \geq 1$. $\square$

Although the assumption in the above theorem, may seem to be very restrictive, it is satisfied in many examples including the $n$-cycle (see Subsection 2.3.2). The method of the proof on the other hand, helps us generalize the result for a large class of graphs including certain random graphs.

Recall the definition of graph isomorphism from Subsection 1.2.1. Following Aldous and Steele (2004), we say a sequence of rooted random or deterministic graphs $\{(G_n, v_0^n)\}_{n\geq 1}$ with roots $\{v_0^n\}_{n\geq 1}$ converges to a random or deterministic graph $(G_\infty, v_0^\infty)$ in the sense of *local weak convergence (l.w.c)* and write $(G_n, v_0^n) \xrightarrow{l.w.c.} (G_\infty, v_0^\infty)$ if for any $d \geq 1$,

$$\mathbb{P}\left(N_d(G_n, v_0^n) \cong N_d(G_\infty, v_0^\infty)\right) \longrightarrow 1 \text{ as } n \to \infty\,. \tag{2.2.8}$$

where for two rooted graphs $A$ and $B$ with roots $\rho_A$ and $\rho_B$ we say $A \cong B$ and read "$A$ and $B$ are *isomorphic* as *rooted* graphs" if $A$ and $B$ are isomorphic as graphs and the isomorphism sends the root $\rho_A$ to the root $\rho_B$. Note that for a sequence of deterministic graphs, (2.2.8) means that the event occurs for "large" enough $n$.

**Theorem 2.2.5.** *Let $\{(G_n, v_0^n)\}_{n\geq 1}$ be a sequence of rooted deterministic or random graphs with deterministic or randomly chosen roots $\{v_0^n\}_{n\geq 1}$. Suppose that for each $G_n$ the maximum degrees of a vertex is bounded by a fixed constant, namely $\Delta$. Suppose there is a rooted deterministic or random tree $\mathfrak{T}$ with root $\rho$ such that*

$$(G_n, v_0^n) \xrightarrow{l.w.c.} (\mathfrak{T}, \rho) \text{ as } n \to \infty\,. \tag{2.2.9}$$

*Let $LB^{G_n, \{v_0^n\}} := \mathbb{E}\left[Y^{\mathcal{T}_n, \{v_0^n\}}\right]$ where $\mathcal{T}_n$ is a BFS spanning tree rooted at $v_0^n$ of the graph*

$G_n$.

*Then for* $\beta < \frac{1}{\Delta}$

$$\left(\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] - LB^{G_n,\{v_0^n\}}\right) \longrightarrow 0 \ \ as \ \ n \to \infty \,. \tag{2.2.10}$$

*Moreover for* $\beta < \frac{1}{\Delta}$ *we have*

$$\lim_{n\to\infty} LB^{G_n,\{v_0^n\}} = \lim_{n\to\infty} \mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] = \mathbb{E}\left[Y^{\mathcal{T},\{\rho\}}\right] \,. \tag{2.2.11}$$

*Proof.* Let $\mathcal{T}_n$ be a BFS spanning tree rooted at $v_0^n$ of the graph $G_n$ and also as defined earlier let $\mathrm{LB}^{G_n,\{v_0^n\}} = \mathbb{E}\left[Y^{\mathcal{T}_n,\{v_0^n\}}\right]$. Fix $d \geq 1$ and $E_n$ be the event $[N_d\left(G_n, v_0^n\right) \cong N_d\left(\mathcal{T}, \rho\right)]$. Therefore from Theorem 2.2.3

$$\mathrm{LB}^{G_n,\{v_0^n\}} \leq \mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] = \mathbb{E}\left[Y^{G_n,\{v_0^n\}}\mathbf{1}_{E_n}\right] + \mathbb{E}\left[Y^{G_n,\{v_0^n\}}\mathbf{1}_{E_n^c}\right] \,. \tag{2.2.12}$$

Now under our assumption, the degree of any vertex of $G_n$ is bounded by $\Delta$ and $\beta < \frac{1}{\Delta}$, so using inequality (2.2.2), we get

$$\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\mathbf{1}_{E_n^c}\right] \leq \frac{1}{1-\beta\Delta}\mathbb{P}\left(E_n^c\right) \,. \tag{2.2.13}$$

Further note that if $E_n$ occurs, $N_d\left(G_n, v_0^n\right)$ is a tree rooted at $v_0^n$ and thus on $E_n$, $N_d\left(G_n, v_0^n\right)$ is a sub-tree of $\mathcal{T}_n$. So

$$Y^{N_d\left(\mathcal{T}_n,v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n} \leq Y^{\mathcal{T}_n,\{v_0^n\}}\mathbf{1}_{E_n} \,.$$

Denote $\partial_d^* N_d\left(\mathcal{T}_n, v_0^n\right)$ the set of infected vertices in $\mathcal{T}_n$ after $d$ units of time starting with one infected vertex $v_0^n$. Hence we have

$$\begin{aligned}
\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\mathbf{1}_{E_n}\right] &\leq \mathbb{E}\left[Y^{N_d\left(\mathcal{T}_n,v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n}\right] + \mathbb{E}\left[Y^{G_n,\partial_d^* N_d\left(\mathcal{T}_n,v_0^n\right)}\mathbf{1}_{E_n}\right] \\
&\leq \mathbb{E}\left[Y^{N_d\left(\mathcal{T}_n,v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n}\right] + \mathbb{E}\left[Y^{G_n,\partial_d^* N_d\left(\mathcal{T}_n,v_0^n\right)}\right] \\
&= \mathbb{E}\left[Y^{N_d\left(\mathcal{T}_{nn},v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n}\right] + \mathbb{E}\left[\mathbb{E}\left[Y^{G_n,\partial_d^* N_d\left(\mathcal{T}_n,v_0^n\right)} \mid \partial_d^* N_d\left(\mathcal{T}_n, v_0^n\right)\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[Y^{N_d\left(\mathcal{T}_n, v_0^n\right), \{v_0^n\}} \mathbf{1}_{E_n}\right] + \frac{1}{1-\beta\Delta} \mathbb{E}\left[\left|\partial_d^* N_d\left(\mathcal{T}_n, v_0^n\right)\right|\right] \\
&\leq \mathrm{LB}^{G_n, \{v_0^n\}} + \frac{(\beta\Delta)^d}{1-\beta\Delta}, \qquad (2.2.14)
\end{aligned}
$$

For the fourth inequality, we use inequality (2.2.2). In the last inequalities, note that there are at most $\Delta^d$ paths of length $d$ from $v_0^n$ and each path has probability $\beta^d$ of infections occurring all along the path. Therefore $\mathbb{E}\left[\left|\partial_d^* N_d\left(\mathcal{T}_n, v_0^n\right)\right|\right] \leq (\beta\Delta)^d$.

So finally combining (2.2.12), (2.2.14) and (2.2.13) we get that for $\beta < \frac{1}{\Delta}$ and for any $d \geq 1$ we have

$$
\left(\mathbb{E}\left[Y^{G_n, \{v_0^n\}}\right] - \mathrm{LB}^{G_n, \{v_0^n\}}\right) \leq \frac{(\beta\Delta)^d}{1-\beta\Delta} + \frac{1}{1-\beta\Delta} \mathbb{P}\left(E_n^c\right). \qquad (2.2.15)
$$

Now under assumption (2.2.9), we have $\lim_{n\to\infty} \mathbb{P}\left(E_n^c\right) = 0$ so we conclude that for any $d \geq 1$

$$
\limsup_{n\to\infty} \left(\mathbb{E}\left[Y^{G_n, \{v_0^n\}}\right] - \mathrm{LB}^{G_n, \{v_0^n\}}\right) \leq \frac{(\beta\Delta)^d}{1-\beta\Delta}. \qquad (2.2.16)
$$

This proves (2.2.10) by taking $d \to \infty$ as $\beta < \frac{1}{\Delta}$.

Now for proving (2.2.11), we first observe that from (2.2.9) the degree of any vertex of $\mathcal{T}$ is also bounded by $\Delta$. So using (2.2.2), we get that for $\beta < \frac{1}{\Delta}$

$$
\mathbb{E}\left[Y^{N_d(\mathcal{T}, \rho), \{\rho\}}\right] \leq \frac{1}{1-\beta\Delta}.
$$

Moreover from the definition, $Y^{N_d(\mathcal{T}, \rho), \{\rho\}} \uparrow Y^{\mathcal{T}, \{\rho\}}$ as $d \to \infty$. So by the Monotone Convergence Theorem we have

$$
\lim_{d\to\infty} \mathbb{E}\left[Y^{N_d(\mathcal{T}, \rho), \{\rho\}}\right] = \mathbb{E}\left[Y^{\mathcal{T}, \{\rho\}}\right] \leq \frac{1}{1-\beta\Delta} < \infty. \qquad (2.2.17)
$$

Thus for fixed $\epsilon > 0$ we can find $d \geq 1$ such that

$$
\left|\mathbb{E}\left[Y^{\mathcal{T}, \{\rho\}}\right] - \mathbb{E}\left[Y^{N_d(\mathcal{T}, \rho), \{\rho\}}\right]\right| < \epsilon \qquad (2.2.18)
$$

and

$$\frac{(\beta\Delta)^d}{1-\beta\Delta} < \epsilon\,. \tag{2.2.19}$$

The last inequality holds as $\beta < \frac{1}{\Delta}$. Further, as degree of any vertex of $\mathcal{T}$ is bounded by $\Delta$ so arguing similar to the derivation of the equation (2.2.13) we conclude

$$\mathbb{E}\left[Y^{N_d(\mathcal{T},\rho),\{\rho\}}\right] - \mathbb{E}\left[Y^{N_d(\mathcal{T},\rho),\{\rho\}}\mathbf{1}_{E_n}\right] = \mathbb{E}\left[Y^{N_d(\mathcal{T},\rho),\{\rho\}}\mathbf{1}_{E_n^c}\right] \leq \frac{1}{1-\beta\Delta}\mathbb{P}\left(E_n^c\right)\,. \tag{2.2.20}$$

Also, arguing similar to the derivation of the equation (2.2.15) we have

$$\begin{aligned}\left|\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] - \mathbb{E}\left[Y^{N_d\left(G_n,v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n}\right]\right| &\leq \frac{(\beta\Delta)^d}{1-\beta\Delta} + \frac{1}{1-\beta\Delta}\mathbb{P}\left(E_n^c\right) \\ &\leq \epsilon + \frac{1}{1-\beta\Delta}\mathbb{P}\left(E_n^c\right)\,, \end{aligned} \tag{2.2.21}$$

where the last equality follows from (2.2.19). Finally,

$$\begin{aligned}\left|\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] - \mathbb{E}\left[Y^{\mathcal{T},\{\rho\}}\right]\right| &\leq \left|\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] - \mathbb{E}\left[Y^{N_d\left(G_n,v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n}\right]\right| \\ &\quad + \left|\mathbb{E}\left[Y^{N_d\left(G_n,v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n}\right] - \mathbb{E}\left[Y^{N_d(\mathcal{T},\rho),\{\rho\}}\right]\right| \\ &\quad + \left|\mathbb{E}\left[Y^{N_d(\mathcal{T},\rho),\{\rho\}}\right] - \mathbb{E}\left[Y^{\mathcal{T},\{\rho\}}\right]\right| \\ &\leq 2\epsilon + \frac{2}{1-\beta\Delta}\mathbb{P}\left(E_n^c\right)\,,\end{aligned}$$

where the last inequality follows from the equations (2.2.18), (2.2.19), (2.2.20) and (2.2.21) and also observing the fact that $\mathbb{E}\left[Y^{N_d\left(G_n,v_0^n\right),\{v_0^n\}}\mathbf{1}_{E_n}\right] = \mathbb{E}\left[Y^{N_d(\mathcal{T},\rho),\{\rho\}}\mathbf{1}_{E_n}\right]$. Now under our assumption (2.2.9) we have $\mathbb{P}\left(E_n\right) \longrightarrow 1$. So we conclude that

$$\lim_{n\to\infty}\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] = \mathbb{E}\left[Y^{\mathcal{T},\{\rho\}}\right]\,. \tag{2.2.22}$$

Thus using (2.2.10), it follows that

$$\lim_{n\to\infty}\mathbf{LB}^{G_n,\{v_0^n\}} = \lim_{n\to\infty}\mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] = \mathbb{E}\left[Y^{\mathcal{T},\{\rho\}}\right]\,.$$

This completes the proof. □

An immediate and interesting application of the above theorem is the following result which gives an explicit formula for the limit of epidemic spread on a randomly selected $r$-regular graph when the infection starts from a randomly chosen vertex.

**Theorem 2.2.6.** *Suppose $G_n$ is a graph, selected uniformly at random from the set of all $r$-regular graphs on $n$ vertices where we assume $nr$ is an even number. Let $v_0^n$ be an uniformly selected vertex of $G_n$. Then for $\beta < \frac{1}{r}$*

$$\lim_{n\to\infty} \mathbb{E}\left[Y^{G_n,\{v_0^n\}}\right] = \frac{1+\beta}{1-(r-1)\,\beta}\,. \tag{2.2.23}$$

We note that in this case, the upper bound given in (Draief et al., 2008) is $\frac{1}{1-r\beta}$ when $\beta < \frac{1}{r}$ which is strictly bigger than the exact answer given in (2.2.23).

*Proof.* It is known (Aldous and Steele, 2004, Janson et al., 2000) that if $G_n$ is a graph selected uniformly at random from the set of all $r$-regular graphs on $n$ vertices, where $nr$ is even and $v_0^n$ be a randomly selected vertex of $G_n$ then

$$(G_n, v_0^n) \xrightarrow{l.w.c.} (\mathbb{T}_r, \rho)\ , \tag{2.2.24}$$

where $\mathbb{T}_r$ is the infinite $r$-regular tree with root say $\rho$. The result then follows from Theorem 2.2.5 and equation (2.3.4). □

### 2.2.2 Starting with more than one infected vertex

Now suppose instead of one infected vertex, we start with $k$ infected vertices which are given by the set $I := \{v_{0,1}, v_{0,2}, \ldots, v_{0,k}\}$. The following theorem gives a lower bound similar to that of Theorem 2.2.3.

**Theorem 2.2.7.** *Let $G$ be an arbitrary finite graph and $I := \{v_{0,j}\}_{j=1}^{k}$ be a fixed set of $k$ vertices. Let $T$ be a spanning forest of the connected components of $G$ containing the vertices in $I$ with* *exactly $k$ trees which are rooted at the vertices in $I$. Then*

$$\mathbb{E}\left[Y^{T,I}\right] \leq \mathbb{E}\left[Y^{G,I}\right] \;\; \textit{for all} \;\; 0 < \beta < 1\,. \tag{2.2.25}$$

*Moreover, if $\mathcal{T}$ is a* breath-first-search spanning forest *of the connected components of $G$ containing the vertices in $I$ with exactly $k$ trees which are rooted at the vertices in $I$ then*

$$\mathbb{E}\left[Y^{T,I}\right] \leq \mathbb{E}\left[Y^{\mathcal{T},I}\right] \leq \mathbb{E}\left[Y^{G,I}\right] \;\; \textit{for all} \;\; 0 < \beta < 1\,. \tag{2.2.26}$$

Given a finite labeled graph $G$ and a fixed set of vertices $I = \{v_{0,j}\}_{j=1}^{k}$ of it, by a *breath-first-search spanning forest* of the connected components of $G$ containing the vertices in $I$ with exactly $k$ trees which are rooted at the vertices in $I$, we mean a spanning forest of $G$ with exactly $k$ connected components which are rooted at the vertices $\{v_{0,1}, v_{0,2}, \cdots, v_{0,k}\}$, that are obtained through the *breath-first-search* algorithm, starting at some vertex $v \in I$ and assuming that all the vertices $\{v_{0,1}, v_{0,2}, \cdots, v_{0,k}\}$ are at the same level. Alternately, we can consider a new graph $G^*$ which is same as $G$ except it has one "artificial" vertex, say $v^*$ which is connected to the vertices $v_{0,1}, v_{0,2}, \cdots, v_{0,k}$ through $k$ "artificial" edges and we perform the BFS algorithm on $G^*$ starting with the vertex $v^*$, to obtain a BFS spanning tree, say $\mathcal{T}^*$ of $G^*$ rooted at $v^*$. Then a *breath-first-search spanning forest* of $G$ with exactly $k$ trees which are rooted at the vertices $\{v_{0,1}, v_{0,2}, \cdots, v_{0,k}\}$ is given by the forest $\mathcal{T}^* \setminus \{v^*\}$. This alternate description, is quite useful in practice. Note that if $\{\mathcal{T}_i\}_{1 \leq i \leq k}$ are the $k$ connected components, rooted respectively at $\{v_{0,1}, v_{0,2}, \cdots, v_{0,k}\}$ of $\mathcal{T}$, a breath-first-search spanning forest of the connected components of $G$ containing the vertices in $I$, then the following identity holds for every $\beta \in (0, 1)$ :

$$\mathbb{E}\left[Y^{\mathcal{T},I}\right] = \sum_{i=1}^{k} \mathbb{E}\left[Y^{\mathcal{T}_i,I}\right] = \frac{\mathbb{E}\left[Y^{\mathcal{T}^*,\{v^*\}}\right] - 1}{\beta}\,. \tag{2.2.27}$$

Using the above identity, we can now generalize all the results of the previous section for epidemic spread starting with more than one infected vertex.

We write $\mathrm{LB}^{G,I}$ for $\mathbb{E}\left[Y^{\mathcal{T},I}\right]$ which is the lower bound of $\mathbb{E}\left[Y^{G,I}\right]$ for starting with $k$ infected vertices given by $I$. Observe that from equation (2.2.27) we can write

$$\mathrm{LB}^{G,I} = \sum_{i=1}^{k} \mathbb{E}\left[Y^{\mathcal{T}_i,I}\right] , \tag{2.2.28}$$

where $\mathcal{T} = \overset{k}{\underset{i=1}{\cup}} \mathcal{T}_i$ is as above. It is worth nothing here that the lower bound $\mathrm{LB}^{G,I}$ does not depend on the choice of $\mathcal{T}$ but the representation given in equation (2.2.28) uses a specific choice of $\mathcal{T}$.

**Theorem 2.2.8.** *Let $\{(G_n, I_n)\}_{n\geq 1}$ be a sequence of graphs where each $G_n$ has k-roots given by the set $I_n := \left\{v_{0,1}^n, v_{0,2}^n, \cdots, v_{0,k}^n\right\}$ such that there exists a sequence $\alpha_n = \Omega(\log n)$ with $N_{\alpha_n}(G_n, I_n) := \overset{k}{\underset{j=1}{\cup}} N_{\alpha_n}\left(G_n, v_{0,j}^n\right)$ is a forest with k components. Then, there exists $0 < \beta_0 \leq 1$, such that for all $0 < \beta < \beta_0$*

$$\left|\mathbb{E}\left[Y^{G_n,I_n}\right] - LB^{G_n,I_n}\right| \longrightarrow 0 \;\; as \;\; n \to \infty \tag{2.2.29}$$

*and therefore $\frac{\mathbb{E}\left[Y^{G_n,I_n}\right]}{LB^{G_n,I_n}} \longrightarrow 1$ as $n \to \infty$.*

The proof of this result is similar to that of Theorem 2.2.4 and follows from the identity (2.2.27). The details are thus omitted.

Our next result is parallel to the Theorem 2.2.5 which needs a generalization of the concept of local weak convergence which was introduced by Wästlund (2012).

We will say a sequence of random or deterministic graphs $\{G_n\}_{n\geq 1}$ with $k$ roots given by the set $I_n := \left\{v_{0,1}^n, v_{0,2}^n, \cdots, v_{0,k}^n\right\}$, $n \geq 1$ converges to a random or deterministic graph $G_\infty$ with $k$-roots say $I_\infty := \left\{v_{0,1}^\infty, v_{0,2}^\infty, \cdots, v_{0,k}^\infty\right\}$ in the sense of *local weak convergence (l.w.c)* and write $(G_n, I_n) \xrightarrow{l.w.c.} (G_\infty, I_\infty)$ if for any $d \geq 1$

$$\mathbb{P}\left(N_d\left(G_n, v_{0,j}^n\right) \cong N_d\left(G_\infty, v_{0,j}^\infty\right) \;\text{ for all }\; 1 \leq j \leq k\right) \longrightarrow 1 \;\text{ as } n \to \infty\,. \tag{2.2.30}$$

Note that for a sequence of deterministic graphs, (2.2.30) means that the event occurs for “large” enough $n$.

**Theorem 2.2.9.** *Let $(G_n)_{n\geq 1}$ be a sequence of deterministic or random graphs. Suppose each $G_n$ has deterministic or randomly chosen $k$ roots given by $I_n := \left\{v_{0,1}^n, v_{0,2}^n, \cdots, v_{0,k}^n\right\}$ and the maximum degree of each $G_n$ is bounded by a fixed constant, namely $\Delta$. Suppose $\mathcal{T} := \overset{k}{\underset{j=1}{\cup}} \mathcal{T}_j$ is a forest with $k$ rooted tress with roots $I_\infty := \{\rho_1, \rho_2, \cdots, \rho_k\}$. We assume that*

$$(G_n, I_n) \xrightarrow{l.w.c.} (\mathcal{T}, I_\infty) \quad as \ \ n \to \infty\,. \tag{2.2.31}$$

*Then for $\beta < \frac{1}{\Delta}$*

$$\left(\mathbb{E}\left[Y^{G_n,I_n}\right] - LB^{G_n,I_n}\right) \longrightarrow 0\,, \tag{2.2.32}$$

*as $n \to \infty$. Moreover*

$$\lim_{n\to\infty} LB^{G_n,I_n} = \lim_{n\to\infty} \mathbb{E}\left[Y^{G_n,I_n}\right] = \mathbb{E}\left[Y^{\mathcal{T},I_\infty}\right] = \sum_{j=1}^{k} \mathbb{E}\left[Y^{\mathcal{T}_j,\{\rho_j\}}\right]\,. \tag{2.2.33}$$

*Proof.* For each $n \geq 1$ as done above we define a new rooted graph $G_n^*$ with artificial vertex $v_n^*$ which is connected to the $k$-roots in $I_n$ of $G_n$ through $k$ artificial edges. Also we consider $\mathcal{T}^*$ defined similarly with an artificial root $\rho^*$ connecting to $\{\rho_1, \rho_2, \cdots, \rho_k\}$. Then our assumption of local weak convergence (2.2.31) is equivalent to

$$(G_n^*, v_n^*) \xrightarrow{l.w.c.} (\mathcal{T}^*, \rho^*) \quad \text{as} \ \ n \to \infty\,. \tag{2.2.34}$$

This together with the relation (2.2.27) and Theorem 2.2.5 completes the proof.

$\square$

It is worth noting that in case $\{\mathcal{T}_j\}_{1\leq j\leq k}$ are i.i.d. (if they are random) or isomorphic (if they are constant), then equation (2.2.33) can be reformulated as

$$\lim_{n\to\infty} \mathrm{LB}^{G_n,I_n} = \lim_{n\to\infty} \mathbb{E}\left[Y^{G_n,I_n}\right] = \mathbb{E}\left[Y^{\mathcal{T},I_\infty}\right] = k\,\mathbb{E}\left[Y^{\mathcal{T}_1,\{\rho_1\}}\right]\,. \tag{2.2.35}$$

As in the case of starting with one infected vertex, the following theorem is an immediate application of the above results.

**Theorem 2.2.10.** *Suppose $G_n$ is a graph, selected uniformly at random from the set of all $r$-regular graphs on $n$ vertices where we assume $nr$ is an even number. Let $I := \left\{v_{0,j}^n\right\}_{j=1}^k$ be $k$ uniformly and independently selected vertices of $G_n$. Then for $\beta < \frac{1}{r}$*

$$\lim_{n\to\infty} \mathbb{E}\left[Y^{G_n,I_n}\right] = k\frac{1+\beta}{1-(r-1)\beta}\,. \tag{2.2.36}$$

*Proof.* Since the vertices in $I_n$ are selected uniformly at random, from Aldous and Steele (2004) we have

$$(G_n, I_n) \xrightarrow{l.w.c.} (\mathcal{T}_r, I_\infty)\ , \tag{2.2.37}$$

where $I_\infty := \{\rho_1, \rho_2, \cdots, \rho_k\}$ and $\mathcal{T}_r$ is a forest with $k$ infinite $r$-regular tree with roots in $I_\infty$. The result then follows from Theorems 2.2.9 and 2.2.6. □

Once again we note that in this case, the upper bound $\frac{k}{1-r\beta}$ given in (Draief et al., 2008) for $\beta < \frac{1}{r}$, is strictly bigger than the exact answer given in (2.2.36) and the gap increases with $k$, the initial number of infections.

## 2.3 Examples

### 2.3.1 Tree

If $G$ is a tree and the epidemic starts with only one infected vertex say $\rho$ which we call the root, then from the construction of the lower bound it is clear that $\mathrm{LB}^{G,\{\rho\}} = \mathbb{E}\left[Y^{G,\{\rho\}}\right]$. In certain cases one can find explicit formula for this quantity. Two such examples are discussed below.

**Regular Tree** Consider a rooted $r$-array tree ($r \geq 2$), with height $m$, denote it by $T(r,m)$. The height of a rooted tree is the length of a longest path from the root. In $T(r,m)$ every internal vertex, except the root $\rho$ has degree $r$. A vertex $v$ is said to be an internal vertex, if it has a neighbor which is not on the unique path from $v$ to $\rho$. We assume that the degree of the root $\rho$ is

$(r-1)$. Figure 2.1 shows a rooted 4-regular tree with height 2.

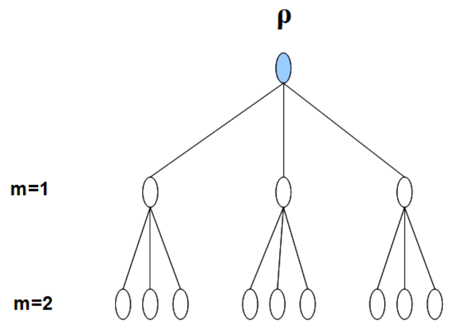


Figure 2.1: $T(4,2)$, rooted 4-regular tree with height 2

Let $\mu_m := \mathbb{E}[Y^{T(r,m),\{\rho\}}]$. Note that the total number of vertices in $T(r,m)$ is $\frac{(r-1)^{m+1}-1}{r-2}$. Now, to calculate the exact value of $\mu_m$ we note that

$$\mu_m = 1 + (r-1)\,\beta\mu_{m-1} \tag{2.3.1}$$

which gives the formula

$$\mu_m = \frac{[(r-1)\,\beta]^{m+1} - 1}{(r-1)\,\beta - 1}\,. \tag{2.3.2}$$

As $T(r,m)$ is a tree, so the lower bound is exact, that is, $\mathrm{LB}^{T(r,m),\{\rho\}} = \mu_m$. Now the upper bound from Draief et al. (2008) is $\frac{1}{1-r\beta}$ for $\beta < \frac{1}{r}$. If $\beta < \frac{1}{r}$ then by Theorem 2.2.5 we get

$$\mathbb{E}\left[Y^{T(r),\{\rho\}}\right] = \lim_{m\to\infty} \mu_m = \frac{1}{1-(r-1)\,\beta}\,, \tag{2.3.3}$$

where $T(r)$ is the rooted infinite $r$-regular tree, where each vertex except the root $\rho$ has degree $r$ and the degree of the root is $(r-1)$.

We observe a gap between the lower bound (which in this case agrees with $\mu_m$ ) to that of the upper bound obtained from Draief et al. (2008).

Now let $\mathbb{T}_r$ be the infinite $r$-regular tree where each vertex including the root, has degree $r$. Such a tree can be viewed as a disjoint union of $r$ rooted infinite $r$-regular trees, whose roots are joined to the root, say $\rho$ of $\mathbb{T}_r$. Thus from (2.3.3) we get that for $\beta < \frac{1}{r}$

$$\mathrm{LB}^{\mathbb{T}_r,\{\rho\}} = \mathbb{E}\left[Y^{\mathbb{T}_r,\{\rho\}}\right] = 1 + \frac{r\beta}{1-(r-1)\,\beta} = \frac{1+\beta}{1-(r-1)\,\beta}\,. \tag{2.3.4}$$

**Galton-Watson Tree** Consider a Galton-Watson branching process starting with one individual. Let the mean of the offspring distribution be $c > 0$. We denote the random tree generated by this process as GW $(c)$ with root $\rho$. Once again, as discussed above, since GW $(c)$ is a tree, therefore $\mathrm{LB}^{\mathrm{GW}(c),\{\rho\}} = \mathbb{E}\left[Y^{\mathrm{GW}(c),\{\rho\}}\right]$. Now in this case, the epidemic process starting with only one infection at $\rho$, is a Galton-Watson branching process starting with one individual as the root and with mean of the new progeny distribution being $\beta c$. So in particular if $\beta < \frac{1}{c}$ then from standard branching process theory $\mathbb{E}\left[Y^{\mathrm{GW}(c),\{\rho\}}\right] < \infty$ and equals $\frac{1}{1-\beta c}$ (Athreya and Ney, 2004).

**Star graph** A *Star graph*, denote by $S_n$, is a graph consisting of a root $\rho$ and $n-1$ leaves, each of which is attached only to the root. For this graph BFS lower bound and the exact value of $\mathbb{E}[Y^{\mathrm{S}_n,\{\rho\}}]$ is:

$$1 + (n-1)\beta$$

Note that upper bound from Draief et al. (2008) is $\frac{1}{1-\sqrt{n-1}\beta}$, for $\sqrt{n-1}\beta < 1$.

### 2.3.2 Cycle

Cycle graph is a graph that consists of a single cycle. We denote the cycle with $n$ vertices by $C_n$. For simplicity, we assume $n$ is odd and then from the BFS algorithm, it is immediate that starting with one infected individual, say at $v_0^n$, we have

$$\mathrm{LB}^{C_n,\{v_0^n\}} = 1 + 2\left(\beta + \beta^2 + \cdots + \beta^{\frac{n-1}{2}}\right) \tag{2.3.5}$$

which converges to $\frac{1+\beta}{1-\beta}$ as $n \to \infty$ for any $0 < \beta < 1$. Now it is clear from the definition that

$$(C_n, v_0^n) \xrightarrow{l.w.c.} (\mathbb{Z}, 0)\ . \tag{2.3.6}$$

Thus using Theorem 2.2.5 we conclude that if $\beta < \frac{1}{2}$ then

$$\lim_{n\to\infty} \mathrm{LB}^{C_n,\{v_0^n\}} = \lim_{n\to\infty} \mathbb{E}\left[Y^{C_n,\{v_0^n\}}\right] = \frac{1+\beta}{1-\beta}. \tag{2.3.7}$$

In fact this holds for any $0 < \beta < 1$. This is because for a cycle graph, the assumption in Theorem 2.2.4 holds for $\alpha_n = n/3$. Thus from the proof of Theorem 2.2.4, we conclude that the equation (2.3.7) holds for any $0 < \beta < 1$.

Now if the epidemic starts with $k$ initial infected vertices given by $I_n := \left\{v_{0,1}^n, v_{0,2}^n, \cdots, v_{0,k}^n\right\}$ which are uniformly distributed, then it is easy to see that

$$(C_n, I_n) \xrightarrow{l.w.c.} (\mathbb{Z}_j, 0)_{1\le j\le k}, \tag{2.3.8}$$

where $\mathbb{Z}_j$ is just a copy of $\mathbb{Z}$. Then by Theorem 2.2.9 we conclude that for $0 < \beta < \frac{1}{2}$,

$$\lim_{n\to\infty} \mathrm{LB}^{C_n,I_n} = \lim_{n\to\infty} \mathbb{E}\left[Y^{C_n,I_n}\right] = k\,\frac{1+\beta}{1-\beta}. \tag{2.3.9}$$

As earlier, we can use Theorem 2.2.8 with $\alpha_n = \Omega(n)$ to conclude that (2.3.9) holds for all all $0 < \beta < 1$.

### 2.3.3 Generalized Cycle

Suppose in a cycle graph, we choose randomly without replacement, $2m$ vertices and connect these vertices by joining edges between them, where $m \ge 1$ is fixed. We call this graph a *Generalized Cycle* and denote it by $\mathrm{GC}(n, m)$. Now consider the epidemic model on this graph with one initial infected vertex $v_0^n$. For large enough $n$, the probability of having at least one of the $m$ pairs inside a neighborhood of $v_0^n$ of radius $r$ is given by

$$1 - \left(1 - \frac{2r(2r+1)}{n(n-1)}\right)^m$$

which tends to zero as $n \to \infty$. Therefore, a fixed neighborhood of the root is a tree with high probability, in fact it is isomorphic to a neighborhood of integer line. Hence by Theorem 2.2.5 it follows that for $\beta < \frac{1}{2}$

$$\lim_{n\to\infty} \mathrm{LB}^{\mathrm{GC}(n,m),\{v_0^n\}} = \lim_{n\to\infty} \mathbb{E}\left[Y^{\mathrm{GC}(n,m),\{v_0^n\}}\right] = \frac{1+\beta}{1-\beta}\,. \tag{2.3.10}$$

Similarly if we start with $k$ initial infected vertices, say $I_n := \left\{v_{0,j}^n\right\}_{j=1}^k$ which are chosen uniformly at random, then it is easy to see that

$$\left(\mathrm{GC}\left(n,m\right), I_n\right) \xrightarrow{l.w.c.} \left(\mathbb{Z}_j, 0\right)_{1\leq j\leq k}\,, \tag{2.3.11}$$

where $\mathbb{Z}_j$ is just a copy of $\mathbb{Z}$. Thus by Theorem 2.2.9 we get

$$\lim_{n\to\infty} \mathrm{LB}^{\mathrm{GC}(n,m),I_n} = \lim_{n\to\infty} \mathbb{E}\left[Y^{\mathrm{GC}(n,m),I_n}\right] = k\frac{1+\beta}{1-\beta}\,, \tag{2.3.12}$$

when $\beta < \frac{1}{3}$, because the maximum degree in $\mathrm{GC}\left(n,m\right)$ is 3.

### 2.3.4 Cube graph

The cube graph is the graph obtained from the vertices and edges of the 3-dimensional unit cube. We denote it by $Q_3$. Suppose initially only the vertex $(0,0,0)$ is infected. Consider a BFS spanning tree $\mathcal{T}$ of $Q_3$ rooted at $(0,0,0)$. Since $Q_3$ has only $8$ vertices so $Y^{\mathcal{T},\{(0,0,0)\}}$ takes values $\{0,1,2,3,4,5,6,7\}$ and

$$\begin{aligned}
\mathrm{LB}^{\mathcal{T},\{(0,0,0)\}} &= \mathbb{E}\left[Y^{\mathcal{T},\{(0,0,0)\}}\right] \\
&= 1+3\beta+3\beta^2+\beta^3 \\
&= \left(1+\beta\right)^3\,.
\end{aligned}$$

Figure 2.2 shows how to obtain the BFS spanning tree on Cube graph.

In general, the $d$-dimensional cube graph say $Q_d$ is a $d$-regular graph which has $n = 2^d$ vertices. Following a similar calculation as done above, one can show that for an epidemic

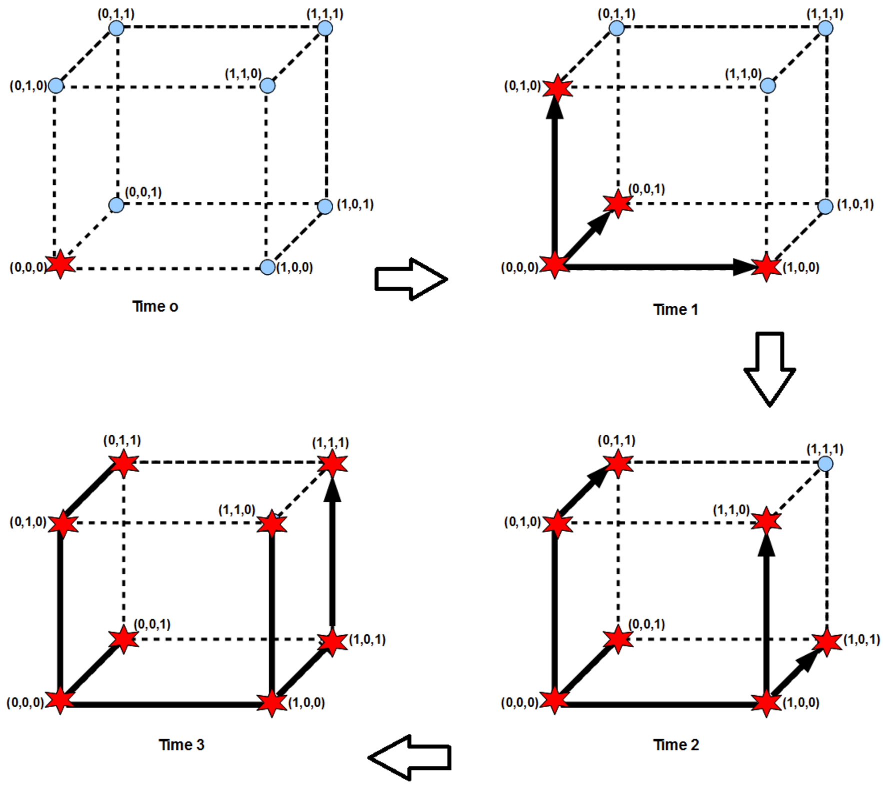


Figure 2.2: BFS spanning tree on Cube graph

starting at one vertex, the lower bound obtained in Theorem 2.2.3 for the expected total number of vertices ever infected is given by $(1+\beta)^d$.

In this example, computation of the exact value of $\mathbb{E}\left[Y^{Q_d,\{(0,0,0)\}}\right]$ is difficult, but we note that there is a gap between the upper bound in (2.2.2), namely $\frac{1}{1-d\beta}$ which is valid only when $\beta < \frac{1}{d}$ and our lower bound. However this is an example which does not fall under any of the theorem we discussed in this chapter and hence we are not sure if the lower bound gives a good approximation.

## 2.4 Discussion

The goal of this study has been to get a better idea of the expected total number of vertices ever infected with as little assumption as possible on the underlying graph $G$. Our approach has been to find an appropriate lower bound of this expectation. Although from a practical point of view, approximation from above with an upper bound is a more conservative method. As shown in the examples given in Section 2.3, the only known upper bounds obtained in Draief et al. (2008) often over estimate the exact quantity. Moreover the upper bounds in Draief et al. (2008) hold only for "small" values of the parameter $\beta$. For an arbitrary finite network, we have obtained a lower bound of the expectation of the number of vertices ever infected for any value of the parameter $\beta$ which is computable through the breadth-first search algorithm. Theorems 2.2.4, 2.2.5, 2.2.8 and 2.2.9 show that this lower bound is asymptotically exact for a large class of graphs when $\beta$ value is "small", which always includes the values of $\beta$ for which the upper bounds from Draief et al. (2008) hold.

However, we would also like to mention here that even though the lower bound we present, works for any infection parameter $0 < \beta < 1$, if the underlying graph has many loops, such as the complete graph $K_n$, then it does not necessarily give a good approximation. To see this, consider the complete graph $K_n$ and suppose that the epidemic starts at a fixed vertex $v_0$. Then the lower bound $\text{LB}^{K_n,\{v_0^n\}} = 1 + (n-1)\,\beta$. Now, let $X_1$ be the number of infected vertices at time $t = 1$. In this case it is easy to see that $X_1 \sim \text{Binomial}\,(n-1, \beta)$. Let $u$ be one of $n-1-X_1$ vertices which are not infected at time $t = 1$. Since $K_n$ is the complete graph, so the conditional probability of $u$ becomes infected at time $t = 2$ given $X_1$ is $1-(1-\beta)^{X_1}$. Hence

$$\begin{aligned}
\mathbb{E}\left[Y^{K_n,\{v_0\}}\right] &\geq 1 + (n-1)\,\beta + \mathbb{E}\left[(n-1-X_1)\left(1-(1-\beta)^{X_1}\right)\right] \\
&= 1 + (n-1)\,\beta + (n-1) - (n-1)\left(1-\beta^2\right)^{n-1} \\
&\qquad - (n-1)\,\beta + (n-1)\,\beta\,(1-\beta)\left(1-\beta^2\right)^{n-2}
\end{aligned}$$

Therefore we get

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[Y^{K_n,\{v_0\}}\right] - \text{LB}^{K_n,\{v_0^n\}}}{\text{LB}^{K_n,\{v_0^n\}}} \geq \frac{1-\beta}{\beta}\,. \tag{2.4.1}$$

where $\text{LB}^{K_n,\{v_0^n\}} := \mathbb{E}[Y^{\mathcal{T}_n,\{v_0^n\}}]$. Here, it is worth mentioning that for the complete graph if we start with one infected vertex, then as discussed in Section 2.1, the set of vertices ever infected is no other than an Erdös-Rényi random graph with parameter $n$ and $\beta$. Thus asymptotic behavior of $\mathbb{E}\left[Y^{K_n,\{v_0\}}\right]$ is well understood in the literature (Bollobás, 2001, Janson et al., 2000).

# Chapter 3

# Nearest neighbor algorithm for the mean field TSP [1]

## 3.1 Introduction

The traveling salesman problem (TSP) is a very well known combinatorial optimization problem. The aim is to find the shortest tour, connecting a number of cities visited by a traveling salesman on his sales route, such that he visits each city exactly once and finally returns to the starting city. Formally, we are given a set $\{c_1, c_2, \ldots, c_n\}$ of *cities* and for each pair $\{c_i, c_j\}$ of distinct cities, a distance $d(c_i, c_j)$. The goal is to find a permutation $\pi$ of the cities that minimizes the quantity

$$\sum_{i=1}^{n} d(c_{\pi(i)}, c_{\pi(i+1)}) \tag{3.1.1}$$

where $\pi(n+1) = 1$. This quantity is called the *tour length*, since it is the total distance traveled by the salesman. We shall concentrate in this chapter on the *symmetric* TSP, in which the distances satisfy

$$d(c_i, c_j) = d(c_j, c_i) \quad \text{for} \quad 1 \leq i, j \leq n.$$

There are several randomized versions of this problem where the distances are taken to be

[1] This chapter is based on the paper by Bandyopadhyay and Sajadi (2013)

random. In particular the one which attracted considerable attention among mathematicians and computer scientists is known as the *Euclidean TSP*, in which the $n$ cities are randomly distributed in a $d$-dimensional hypercube and the distances between cities are given by the Euclidean metric and are thus random. The other random TSP, which has been of interest within the statistical physics community is *the mean field TSP*. Here the distances between pairs of cities, i.e., $d(c_i, c_j)$ are taken as independent random variables with a given distribution $F$. Note that in this case, the geometric structure may break since the triangle inequality may not necessarily hold with probability one. In fact we cannot quite say that the numbers $d(c_i, c_j)$ really represent distances under any metric. Although this seems artificial, however such models are of interest in statistical physics literature.

It is well known in algorithm literature (Papadimitriou and Steiglitz, 1998) that TSP in general is a *NP-Complete* problem. So there are several approximate algorithms which tries to approximate the optimal tour with polynomial running time. Among them, one of the simplest is the *Nearest Neighbor (NN) Algorithm*, which has been described in Subsection 1.2.2.

Denote the distance $d(c_i, c_j)$ by $L_{ij}$. Since the NN algorithm is to move to the nearest non-visited city, therefore starting from $c_1$, by using this algorithm we need to find the nearest city to it. We call it $v_2$. In this way, we need to find

$$\min\{L_{12}, L_{13}, \ldots, L_{1n}\}$$

Then from city $v_2$ we find the nearest city to that and call it $v_3$. Here we need to find

$$\min\{L_{v_2 u} | u \in \{2, 3, \ldots, n\} \quad \text{and} \quad u \neq v_2\}.$$

We continue the algorithm till all $n$ cities have been visited. Then from there we go back to starting city which is $c_1$.

Define $T_n^{NN}$ to be the length of NN tour among $n$ cities in the TSP, then

$$T_n^{NN} = \sum_{i=1}^{n} L_{v_i v_{i+1}}, \quad v_1 = 1 = v_{n+1} \tag{3.1.2}$$

### 3.1.1 The deterministic TSP

The performance of nearest neighbor algorithm has been studied for the TSP when the distances are defined through a metric. Let $T_n^{opt}$ be the length of the optimal tour and $\lceil x \rceil$ denote the smallest integer greater than or equal to $x$. Rosenkrantz et al. (1977) measured the closeness of a tour by the ratio of the obtained tour length, to the optimal tour length. They proved that if the cities are placed in a metric space and the intercity distances are given by the metric then

$$\frac{T_n^{NN}}{T_n^{opt}} \leq \frac{1}{2} \lceil \log_2 n \rceil + \frac{1}{2} .$$

They also showed that for each $m > 3$, there exists a traveling salesman graph with $n = 2^m - 1$ nodes inside a metric space such that

$$\frac{T_n^{NN}}{T_n^{opt}} > \frac{1}{3} \log_2 (n+1) + \frac{4}{9} .$$

### 3.1.2 The random TSP

One of the famous mathematical results for the Euclidean TSP is Beardwood-Halton-Hammersley theorem which studies the large sample behavior of the length of shortest tour in TSP. Let the cities be independently and uniformly distributed on $[0, 1]^d$. Beardwood et al. (1959) showed that there is a constant $0 < \beta_{TSP}(d) < \infty$ such that with probability one

$$\frac{T_n^{opt}}{n^{\frac{d-1}{d}}} \longrightarrow \beta_{TSP}(d)$$

They also proved that for nonuniform random samples, there is an universal constant $\beta_{TSP}(d)$ such that

$$\frac{T_n^{opt}}{n^{\frac{d-1}{d}}} \longrightarrow \beta_{TSP}(d) \int_{\mathbb{R}^d} f(x)^{(d-1)/d} dx \quad a.s.$$

where $f(x)$ is the density of the absolutely continuous part of the distribution of cities with a compact support.

Asymptotic results in the mean field TSP have been obtained by Wästlund (2010). Let $L_{ij}$'s be independent random variables from a fixed distribution on the nonnegative real numbers. Suppose as $t \longrightarrow 0^+$

$$\frac{\mathbb{P}(L_{ij} < t)}{t} \longrightarrow 1$$

He proved that for large $n$,

$$T_n^{opt} \xrightarrow{\mathbb{P}} \frac{1}{2}\int_0^{\infty} h(x)\,dx \tag{3.1.3}$$

where $h$ as a function of $x$ is implicitly defined through the equation

$$\left(1+\frac{x}{2}\right)e^{-x} + \left(1+\frac{h(x)}{2}\right)e^{-h(x)} = 1$$

Although there seems to be no simple expression for this limit in terms of known mathematical constants, it can be evaluated numerically to be approximately $2.041548$.

In this chapter we study the limiting behavior of the total length of the tour, obtained by NN algorithm for the mean field TSP. Our motivation is similar to that of Rosenkrantz et al. (1977). We would like to compare the apparent "loss" (that is, more distance to be traversed) accrued by using the NN algorithm with respect to the optimal solution. But because of (3.1.3), it is enough to consider the limiting behavior of $T_n^{NN}$. We show that if $F$, the distribution of the distance between cities, has a density which is continuous at 0 with $F'(0+) > 0$, then the total length of the NN tour for mean field TSP scales as $\log n$. This parallels the conclusions drawn in Rosenkrantz et al. (1977) for Euclidean TSP. Moreover we also consider a general distribution function $F$ with non-negative support and show that the asymptotic behaviors for $T_n^{NN}$ depend on the limiting properties of the density near 0.

The rest of the chapter is structured as follows. In Section 3.2, we study the last edge of NN tour in the mean field TSP. The main results are presented in Section 3.3. Section 3.4 provides some technical results which we use in the proof of main results. Section 3.5 includes the discussion about the assumptions on distribution $F$ and also the relation of objective function with lower records.

## 3.2 The last edge of the NN tour

We will assume that the mean and the variance of $F$ are finite and $F$ has a density $f$. Let the distances between cities be denoted by $\{L_{ij}\}_{1\leq i\leq j\leq n}$ which are i.i.d with distribution $F$ supported on $[0,\infty)$ with $0\in$ support $(F)$ and density $f$. Let $L_n^{\text{last}}$ be the length of the last edge, which joins the last visited city to the first city. Then the length of NN tour, $T_n^{NN}$, can be written as

$$T_n^{NN} \stackrel{d}{=} \sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij} + L_n^{\text{last}} \tag{3.2.1}$$

Let $L_n^{\text{first}} := \min\limits_{1<j\leq n} L_{1j}$. Then (3.2.1) can be rewritten as,

$$T_n^{NN} \stackrel{d}{=} \sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij} + L_n^{\text{first}} + L_n^{\text{last}} \tag{3.2.2}$$

The following proposition shows that the last edge in NN tour does not play an important role.

**Proposition 3.2.1.** *In the NN tour for mean field TSP, the distribution function of $L_n^{first} + L_n^{last}$ converges to $F$ as $n\longrightarrow\infty$ and $\sum_{i=2}^{n-1}\min_{i<j\leq n} L_{ij}$ is independent of $L_n^{first} + L_n^{last}$. Moreover as $n\longrightarrow\infty$,*

$$\mathbb{E}\left[L_n^{first} + L_n^{last}\right] \longrightarrow \mu$$

*and*

$$\mathbb{E}\left[\left(L_n^{first} + L_n^{last}\right)^2\right] \longrightarrow \mu^2+\sigma^2\,,$$

*where $\mu$ and $\sigma^2$ are the mean and the variance of $F$.*

*Proof.* For $k = 1,2,\ldots,n-1$, let $X_k := L_{1k+1}$ and $X_{(k)}$ be the $k^{th}$ order statistic of $X_1, X_2, \ldots, X_{n-1}$. Note that by assumption $X_k$'s are i.i.d. $F$.

Notice that by construction the successive vertices $1 = v_1, v_2, v_3, \ldots, v_n$ of the tour have the property that for every $2\leq k\leq n$ given $\{v_2, v_3, \cdots, v_{k-1}\}$ the vertex $v_k$ is uniformly distributed on the set $\{1,2,\ldots,n\}\setminus\{1, v_2, v_3, \cdots, v_{k-1}\}$. Thus for every $3\leq k\leq n$ given $v_2$, the vertex $v_k$ is uniformly distributed on the set $\{2,3,\ldots,n\}\setminus\{v_2\}$. So in particular the last vertex of the tour $v_n$ is also uniformly distributed on the set $\{2,3,\ldots,n\}\setminus\{v_2\}$. Hence given

$X_1, X_2, \ldots, X_{n-1}$, the length of the last edge is uniform on $\left\{X_{(2)}, X_{(3)}, \ldots, X_{(n-1)}\right\}$. Now for any bounded continuous function $h$ we have,

$$\begin{aligned}
\mathbb{E}\left[h\left(L_n^{\text{last}}\right)\right] &= \frac{1}{n-2}\sum_{k=2}^{n-1}\mathbb{E}\left[h\left(X_{(k)}\right)\right] \\
&= \frac{1}{n-2}\sum_{k=1}^{n-1}\mathbb{E}\left[h\left(X_{(k)}\right)\right] - \frac{\mathbb{E}\left[h\left(X_{(1)}\right)\right]}{n-2} \\
&= \frac{1}{n-2}\sum_{k=1}^{n-1}\mathbb{E}\left[h\left(X_k\right)\right] - \frac{\mathbb{E}\left[h\left(X_{(1)}\right)\right]}{n-2} \\
&= \frac{n-1}{n-2}\mathbb{E}\left[h\left(X_1\right)\right] - \frac{\mathbb{E}\left[h\left(X_{(1)}\right)\right]}{n-2}.
\end{aligned}$$

Therefore

$$\lim_{n\longrightarrow\infty}\mathbb{E}\left[h\left(L_n^{\text{last}}\right)\right] = \mathbb{E}\left[h\left(X_1\right)\right],$$

for every bounded continuous function $h$, thus the distribution function of $L_n^{\text{last}}$ converges to $F$ as $n \longrightarrow \infty$. Now observe that $L_n^{\text{first}} \longrightarrow 0$ almost surely, so by Slutsky's theorem we have the distribution function of $L_n^{\text{first}} + L_n^{\text{last}}$ converges to $F$ as $n \longrightarrow \infty$.

Now observe that by similar calculations as above

$$\mathbb{E}\left[L_n^{\text{first}} + L_n^{\text{last}}\right] = \frac{n-1}{n-2}\mathbb{E}\left[X_1\right] + \frac{n-3}{n-2}\mathbb{E}\left[X_{(1)}\right] \longrightarrow \mu\,.$$

The last limit follows from the dominated convergence theorem by observing that $X_{(1)} \longrightarrow 0$ almost surely and $0 \leq X_{(1)} \leq X_1$.

Further,

$$\mathbb{E}\left[\left(L_n^{\text{last}}\right)^2\right] = \frac{n-1}{n-2}\mathbb{E}\left[X_1^2\right] - \frac{\mathbb{E}\left[X_{(1)}^2\right]}{n-2} \longrightarrow \mu^2 + \sigma^2\,,$$

and

$$\mathbb{E}\left[\left(L_n^{\text{first}}\right)^2\right] = \mathbb{E}\left[X_{(1)}^2\right] \longrightarrow 0\,.$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[L_n^{\text{first}}L_n^{\text{last}}\right] &= \frac{n-1}{n-2}\mathbb{E}\left[X_{(1)}\bar{X}_{n-1}\right] - \frac{\mathbb{E}\left[X_{(1)}^2\right]}{n-2} \qquad \left[\text{where } \bar{X}_{n-1} := \frac{1}{n-1}\sum_{k=1}^{n-1} X_k\right] \\
&\le \sqrt{\mathbb{E}\left[X_{(1)}^2\right]\,\mathbb{E}\left[\bar{X}_{n-1}^2\right]} - \frac{\mathbb{E}\left[X_{(1)}^2\right]}{n-2} \qquad \text{[using Cauchy-Schwarz inequality]} \\
&= \sqrt{\mathbb{E}\left[X_{(1)}^2\right]\left(\mu^2 + \frac{\sigma^2}{n-1}\right)} - \frac{\mathbb{E}\left[X_{(1)}^2\right]}{n-2} \\
&\longrightarrow 0\,.
\end{aligned}$$

Combining all these we have

$$\mathbb{E}\left[\left(L_n^{\text{first}} + L_n^{\text{last}}\right)^2\right] \longrightarrow \mu^2 + \sigma^2\,.$$

□

## 3.3 Main results

For the distribution function $F$ we define $F^{-1} : (0,1) \to [0,\infty)$ by $F^{-1}(u) := \inf\left\{x \in \mathbb{R} \,\middle|\, F(x) \ge u\right\}$, $0 < u < 1$. It is then a standard fact that $F^{-1}(U) \sim F$ when $U \sim$ Uniform $[0,1]$. We start with a lemma which will give an useful representation of $T_n^{NN}$.

**Lemma 3.3.1.** *Let the distances between cities, $(L_{ij})_{i<j\le n}$ for $i = 1,\dots,n-1$ be i.i.d with $F$ denoting its common distribution function. Define the random variable $W_i := F^{-1}\left(1 - \exp(-\frac{Y_i}{i})\right)$ where $\{Y_i\}_{1\le i\le n-1}$ are i.i.d. Exponential random variable each with mean one. Then*

$$\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij} \overset{d}{=} \sum_{i=1}^{n-2} W_i\,.$$

*Thus*

$$T_n^{NN} \overset{d}{=} \sum_{i=1}^{n-2} W_i + R_n\,, \tag{3.3.1}$$

*where $R_n \overset{d}{=} L_n^{first} + L_n^{last}$ and is independent of $\{W_i\}_{i=1}^{n-2}$.*

*Proof.* Let $(\xi_{ij})_{i<j\leq n}$ be i.i.d. Exponential random variable each with mean one. Then

$$\begin{aligned}\sum_{i=2}^{n-1}\min_{i<j\leq n} L_{ij} &\stackrel{d}{=} \sum_{i=2}^{n-1}\min_{i<j\leq n} F^{-1}(1-e^{-\xi_{ij}}) \\ &\stackrel{d}{=} \sum_{i=2}^{n-1} F^{-1}(1-e^{-\min\limits_{i<j\leq n}\xi_{ij}}) \\ &\stackrel{d}{=} \sum_{i=1}^{n-2} F^{-1}(1-e^{-\frac{Y_i}{i}})\end{aligned}$$

where $Y_i$'s are i.i.d. Exponential random variable each with mean one.

Finally (3.3.1) follows from equation (3.2.2). □

In the proofs of our main results, we primarily study properties of $W_i$ rather than $\min\limits_{i<j\leq n} L_{ij}$. Observe that

$$\mathbb{P}(W_i \leq w) = 1 - \{1 - F(w)\}^i \text{ for } w \geq 0. \tag{3.3.2}$$

**Lemma 3.3.2.** *Assume that $F$ has a density $f$ and as $t \longrightarrow 0+$, $\frac{f(t)}{t^\alpha} \longrightarrow C$, where $C \in (0,\infty)$ is constant and $-1 < \alpha < 1$. Then as $n \longrightarrow \infty$, $\{\sum_{i=1}^{n-2}(W_i - \mathbb{E}[W_i])\}_{n\geq 1}$, converges $a.s.$ and in $\mathcal{L}_2$.*

*Proof.* By assumption as $t \longrightarrow 0+$, $\frac{f(t)}{t^\alpha} \longrightarrow C$, therefore given $\epsilon > 0$, there exists $\delta > 0$, such that for all $0 < t < \delta$, we have

$$(C-\epsilon)t^\alpha < f(t) < (C+\epsilon)t^\alpha\,.$$

Hence for $0 < x < \delta$,

$$\frac{(C-\epsilon)}{1+\alpha}x^{1+\alpha} < F(x) < \frac{(C+\epsilon)}{1+\alpha}x^{1+\alpha}$$

which implies

$$(\frac{1+\alpha}{C+\epsilon})^{\frac{1}{1+\alpha}}x^{\frac{1}{1+\alpha}} < F^{-1}(x) < (\frac{1+\alpha}{C-\epsilon})^{\frac{1}{1+\alpha}}x^{\frac{1}{1+\alpha}}. \tag{3.3.3}$$

Put $\delta_1 := -\ln(1-\delta)$. If $\frac{Y_i}{i} < \delta_1$ (which ensures that $1-\exp(-\frac{Y_i}{i}) < \delta$), then we have

$$W_i\,\mathbf{1}\left[\frac{Y_i}{i} < \delta_1\right] < (\frac{1+\alpha}{C-\epsilon})^{\frac{1}{1+\alpha}} \left(1-\exp(-\frac{Y_i}{i})\right)^{\frac{1}{1+\alpha}}\,\mathbf{1}\left[\frac{Y_i}{i} < \delta_1\right]. \tag{3.3.4}$$

Observe that for $\beta > 0$,

$$\begin{aligned}\mathbb{E}\left[\left(1-\exp(-\frac{Y_i}{i})\right)^{\beta}\right] &= \int_0^{\infty} (1-\exp(-y/i))^{\beta}\exp(-y)dy \\ &= i\int_0^1 u^{\beta}(1-u)^{i-1}du \\ &= \Gamma(1+\beta)\frac{\Gamma(i+1)}{\Gamma(i+1+\beta)} \\ &\leq \Gamma(2+\beta)\frac{1}{(i+1+\beta)^{\beta}}.\end{aligned}$$

The last inequality follows from the Wendel's double inequality (Wendel, 1948), which says for real $x > 0$ and $0 < s < 1$ we have

$$\frac{x}{(x+s)^{1-s}}\Gamma(x) \leq \Gamma(x+s) \leq x^s\Gamma(x) \tag{3.3.5}$$

Therefore

$$\mathbb{E}\left[W_i^2\,\mathbf{1}[\frac{Y_i}{i} < \delta_1]\right] < (\frac{1+\alpha}{C-\epsilon})^{\frac{2}{1+\alpha}}\Gamma\left(2+\frac{2}{1+\alpha}\right)\frac{1}{\left(i+1+\frac{2}{1+\alpha}\right)^{\frac{2}{1+\alpha}}}. \tag{3.3.6}$$

Now as $i \longrightarrow \infty$, $\dfrac{Y_i}{i} \xrightarrow{a.s.} 0$. This follows from the Borel-Cantelli lemma, because for any $\epsilon_0 > 0$, the sequence of probabilities $\mathbb{P}\left(Y_i > \epsilon_0\, i\right) = e^{-\epsilon_0\, i}$ are summable. Define

$$I_0(\omega) := \min\left\{i \mid \frac{Y_j(\omega)}{j} < \delta_1,\ \forall j \geq i\right\}. \tag{3.3.7}$$

Fix $m > 1$, then

$$[I_0 = m] = \left[\frac{Y_i}{i} < \delta_1, \forall i \geq m \quad \text{and} \quad \frac{Y_{m-1}}{m-1} > \delta_1\right].$$

Hence

$$\mathbb{P}(I_0 = m) \le e^{-(m-1)\delta_1}$$

Now,

$$\sum_{i=1}^{\infty} \mathbb{E}[W_i^2] = \sum_{m=1}^{\infty} \mathbb{E}[\sum_{i=1}^{m-1} W_i^2 \, \mathbf{1}(I_0 = m)] + \sum_{m=1}^{\infty} \mathbb{E}[\sum_{i=m}^{\infty} W_i^2 \, \mathbf{1}(I_0 = m)]. \tag{3.3.8}$$

But,

$$\mathbb{E}[\sum_{i=1}^{m-1} W_i^2 \, \mathbf{1}(I_0 = m)] = \mathbb{E}[\sum_{i=1}^{m-2} W_i^2 \, \mathbf{1}(I_0 = m)] + \mathbb{E}[W_{m-1}^2 \mathbf{1}(I_0 = m)]$$

Since $[I_0 = m]$ depends on random variables $Y_{m-1}, Y_m, Y_{m+1}, ...$ therefore for $1 \le i \le m-2$, $W_i$ is independent of $[I_0 = m]$, hence

$$\mathbb{E}[\sum_{i=1}^{m-2} W_i^2 \, \mathbf{1}(I_0 = m)] \le e^{-(m-1)\delta_1} \sum_{i=1}^{m-2} \mathbb{E}[W_i^2].$$

Since $\mathbb{E}[W_i^2]$ is a decreasing sequence, we have

$$\sum_{i=1}^{m-2} \mathbb{E}[W_i^2] \le (m-2)\mathbb{E}[W_1^2].$$

Therefore

$$\mathbb{E}[\sum_{i=1}^{m-2} W_i^2 \, \mathbf{1}(I_0 = m)] \le (m-2)e^{-(m-1)\delta_1}\mathbb{E}[W_1^2]. \tag{3.3.9}$$

By Cauchy-Schwarz Inequality

$$\mathbb{E}[W_{m-1}^2 \mathbf{1}(I_0 = m)]] \le \sqrt{\mathbb{E}[W_{m-1}^4]\mathbb{P}(I_0 = m)}.$$

Now for $m > 4$,

$$\mathbb{E}[W_{m-1}^4] \le \mathbb{E}[W_4^4] \le \mu^4. \tag{3.3.10}$$

Therefore

$$\mathbb{E}[W_{m-1}^2 \mathbf{1}(I_0 = m)]] \le \mu^2 e^{-(m-1)\frac{\delta_1}{2}}. \tag{3.3.11}$$

In the last equality of (3.3.10), we use the fact that for $k$ non-negative random variables $Z_1, Z_2, ..., Z_k$,

$$(\min(Z_1, Z_2, ..., Z_k))^k \leq \prod_{j=1}^{k} Z_j.$$

From (3.3.9) and (3.3.11), we have

$$\sum_{m=1}^{\infty} \mathbb{E}[\sum_{i=1}^{m-1} W_i^2 \, \mathbf{1}(I_0 = m)] < \infty \tag{3.3.12}$$

Now we consider the second term of the equation (3.3.8) and observe

$$\begin{aligned}
\sum_{m=1}^{\infty} \mathbb{E}\left[\sum_{i=m}^{\infty} W_i^2 \mathbf{1}\left(I_0 = m\right)\right] &= \sum_{m=1}^{\infty} \left[\sum_{i=m}^{\infty} \mathbb{E}\left[W_i^2 \mathbf{1}\left(I_0 = m\right)\right]\right] \\
&\leq \sum_{m=1}^{\infty} \left[\sum_{i=m}^{\infty} \mathbb{E}\left[W_i^2 \mathbf{1}\left(\frac{Y_i}{i} < \delta_1\right) \mathbf{1}\left(\frac{Y_{m-1}}{m-1} > \delta_1\right)\right]\right] \\
&\qquad \left(\text{as } [I_0 = m] \subseteq \left[\frac{Y_i}{i} < \delta_1 \text{ and } \frac{Y_{m-1}}{m-1} > \delta_1\right] \text{ for all } i \geq m\right) \\
&= \sum_{m=1}^{\infty} \left[\sum_{i=m}^{\infty} \mathbb{E}\left[W_i^2 \mathbf{1}\left(\frac{Y_i}{i} < \delta_1\right)\right] e^{-(m-1)\delta_1}\right] \\
&\qquad (\text{ as } Y_i \text{ and } Y_{m-1} \text{ are independent for all } i \geq m) \\
&\leq \sum_{m=1}^{\infty} e^{-(m-1)\delta_1} \left[\sum_{i=m}^{\infty} K'_\alpha \frac{1}{\left(i+1+\frac{2}{1+\alpha}\right)^{\frac{2}{1+\alpha}}}\right] \quad (\text{by equation (3.3.6)}) \\
&\leq \sum_{m=1}^{\infty} e^{-(m-1)\delta_1} K'_\alpha \left[\sum_{i=m}^{\infty} \frac{1}{i^{\frac{2}{1+\alpha}}}\right] \\
&\leq \sum_{m=1}^{\infty} K''_\alpha e^{-(m-1)\delta_1} \quad \left[\text{as } \frac{2}{1+\alpha} > 1\right] \\
&< \infty
\end{aligned} \tag{3.3.13}$$

where $K'_\alpha = (\frac{1+\alpha}{C-\epsilon})^{\frac{2}{1+\alpha}} \Gamma\left(2 + \frac{2}{1+\alpha}\right)$ and $K''_\alpha$ is a positive constant. Thus from equations (3.3.12) and (3.3.13) we conclude

$$\sum_{i=1}^{\infty} \mathbb{E}[W_i^2] < \infty \tag{3.3.14}$$

Therefore $\mathsf{Var}[\sum_{i=1}^{n} W_i]$ is bounded for all $n$. This shows that $\sum_{i=1}^{n-2}(W_i - \mathbb{E}[W_i])$ as a martingale converges $a.s.$ and in $\mathcal{L}_2$. $\square$

**Theorem 3.3.1.** *Assume that as $t \longrightarrow 0+$, $\frac{f(t)}{t^\alpha} \longrightarrow C$, where $C \in (0, \infty)$ is constant and $-1 < \alpha < 1$. Then as $n \longrightarrow \infty$,*

$$\{T_n^{NN} - \mathbb{E}[T_n^{NN}]\}_{n\geq 1} \quad \textit{converges weakly.} \tag{3.3.15}$$

*Proof.* From equation (3.2.2) we have

$$T_n^{NN} - \mathbb{E}[T_n^{NN}] \stackrel{d}{=} \sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij} - \mathbb{E}[\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij}] + L_n^{\text{first}} + L_n^{\text{last}} - \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}].$$

But by Lemma 3.3.2 and Lemma 3.3.1, $\left\{\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij} - \mathbb{E}[\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij}]\right\}_{n>1}$ converges in $\mathcal{L}_2$ and hence by Proposition 3.2.1, $\left\{T_n^{NN} - \mathbb{E}[T_n^{NN}]\right\}_{n>1}$ converges weakly. $\square$

The next three results consider three cases of the behavior of $f$ near 0. Theorem 3.3.2 covers the case when $f$ near zero converges to a positive constant. In this case, $T_n^{NN}$ scales as constant times $\log n$. Theorem 3.3.3 and Theorem 3.3.4 consider the cases when $\lim_{t\to 0} f(t)$ is zero and infinity respectively. We use the notation $a_n \sim b_n$ to denote $a_n$ is asymptotically equal to $b_n$, that is, $\lim_{n\longrightarrow\infty} \frac{a_n}{b_n} = 1$.

**Theorem 3.3.2.** *Assume that as $t \longrightarrow 0+$, $f(t) \longrightarrow f(0)$, where $f(0) \in (0, \infty)$. Then as $n \longrightarrow \infty$,*

$$\frac{T_n^{NN}}{\log n} \xrightarrow{\mathbb{P}} \frac{1}{f(0)} \tag{3.3.16}$$

*and*

$$\mathbb{E}[T_n^{NN}] \sim \frac{1}{f(0)} \log n. \tag{3.3.17}$$

*Moreover, convergence in* (3.3.16) *happens in* $\mathcal{L}_2$.

*Proof.* We will show

$$\frac{T_n^{NN}}{\log n} \xrightarrow{\mathcal{L}_2} \frac{1}{f(0)} \quad \text{as} \quad n \longrightarrow \infty \,,$$

which will imply (3.3.16). Now,

$$\begin{aligned}
\mathbb{E}\left[\frac{T_n^{NN}}{\log n} - \frac{1}{f(0)}\right]^2 &= \mathbb{E}\left[\frac{T_n^{NN} - \mathbb{E}[T_n^{NN}]}{\log n} + \frac{\mathbb{E}[T_n^{NN}]}{\log n} - \frac{1}{f(0)}\right]^2 \\
&= \frac{\mathbb{E}\left[\left(\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij} - \mathbb{E}[\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij}]\right)^2\right]}{(\log n)^2} \\
&+ \frac{\mathbb{E}\left[\left(L_n^{\text{first}} + L_n^{\text{last}} - \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}]\right)^2\right]}{(\log n)^2} + \left[\frac{\mathbb{E}[T_n^{NN}]}{\log n} - \frac{1}{f(0)}\right]^2 \\
&= \frac{\mathsf{Var}\left[\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij}\right]}{(\log n)^2} + \frac{\mathsf{Var}\left[L_n^{\text{first}} + L_n^{\text{last}}\right]}{(\log n)^2} \\
&+ \left[\frac{\mathbb{E}[T_n^{NN}]}{\log n} - \frac{1}{f(0)}\right]^2 . \qquad (3.3.18)
\end{aligned}$$

Note that $\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij}$ is independent of $L_n^{\text{last}} + L_n^{\text{first}}$. Now by Lemma 3.3.2, Lemma 3.3.1 and Proposition 3.2.1, the first two terms in equation (3.3.18) converges to zero as $n \longrightarrow \infty$. Convergence to zero of the last term in equation (3.3.18) follows from the following observation. By assumption $f(t) \longrightarrow f(0)$ as $t \longrightarrow 0+$, so using the inequality (3.3.3) when $f(0) = C$ and $\alpha = 0$, we get that as $i \longrightarrow \infty$,

$$\frac{f(0)W_i}{\frac{Y_i}{i}} \longrightarrow 1 \quad a.s.$$

where $Y_i$'s are i.i.d. Exponential random variable each with mean one and $W_i = F^{-1}\left(1 - \exp(-\frac{Y_i}{i})\right)$. Therefore as $n \longrightarrow \infty$

$$\frac{f(0)\sum_{i=1}^{n-2} W_i}{\sum_{i=1}^{n-2} \frac{Y_i}{i}} \longrightarrow 1 \quad a.s.$$

Now, since $\mathsf{Var}\left[\sum_{i=1}^{n-2}\frac{Y_i}{i}\right]$ is bounded for all $n$, therefore by the martingale convergence theorem $\sum_{i=1}^{n-2}\frac{Y_i}{i} - \mathbb{E}\left[\sum_{i=1}^{n-2}\frac{Y_i}{i}\right]$ converges almost surely. But $\mathbb{E}\left[\sum_{i=1}^{n}\frac{Y_i}{i}\right] = \sum_{i=1}^{n}\frac{1}{i} \sim \log n$, thus

$$\frac{f(0)\sum_{i=1}^{n-2} W_i}{\log n} \longrightarrow 1 \quad a.s. \tag{3.3.19}$$

Now by Lemma 3.3.2 and Lemma 3.3.1, $\sum_{i=1}^{n-2} W_i - \mathbb{E}[\sum_{i=1}^{n-2} W_i]$ converges $a.s.$ to a random variable. This observation along with (3.3.19) give

$$\lim_{n\longrightarrow\infty} \frac{\mathbb{E}[\sum_{i=1}^{n-2} W_i]}{\log n} = \frac{1}{f(0)} \tag{3.3.20}$$

and therefore by equation (3.3.1) and Proposition 3.2.1,

$$\lim_{n\longrightarrow\infty} \frac{\mathbb{E}[T_n^{NN}]}{\log n} = \frac{1}{f(0)}$$

This also proves $\mathbb{E}[T_n^{NN}] \sim \frac{1}{f(0)} \log n$. □

When the distribution $F$ is Exponential, the expected value of the length of NN tour among $n$ cities scales as $\log n$. This is a special case of Theorem 3.3.2, when $f(0) = 1$. The following corollary is a consequence of Theorem 3.3.2.

**Corollary 3.3.1.** *In the mean field TSP, suppose $F$ is the Exponential distribution with mean one. Then $T_n^{NN} - \log n$ converges weakly.*

*Proof.* Consider a mean field TSP on $n$ cities $\{1, 2, ..., n\}$, where for each $1 \leq i \leq n-1$, the intercity distances $\{L_{ij}\}_{i<j\leq n}$, are i.i.d. Exponential random variable each with mean one. Starting at city 1, our job is to find the nearest city to it, that means to find $\min_{1<j\leq n} L_{1j}$. Now we have a tour, with 2 cities in it. Finding the next nearest city to the last visited city in this tour, in distribution is the same as finding the *minimum* of $n-3$ independent Exponential random

variables.

Since $\min_{i<j\leq n} L_{ij}$ has an Exponential distribution with mean $\frac{1}{n-i}$, then we have

$$\mathbb{E}[\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij}] = \frac{1}{n-1} + \frac{1}{n-2} + \ldots + \frac{1}{2} + 1 \tag{3.3.21}$$

Since $\mathsf{Var}[\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij}] = \sum_{i=1}^{n-1} \frac{1}{i^2}$, hence for all $n \geq 1$, $\mathsf{Var}\left(\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij} - \mathbb{E}[\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij}]\right)$ is bounded. Therefore by the martingale convergence theorem, we conclude that the martingale sequence

$$\left\{\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij} - \mathbb{E}[\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij}]\right\}_{n\geq 1} \quad \text{converges} \quad a.s. \quad \text{and in} \quad \mathcal{L}_2. \tag{3.3.22}$$

Note that as we saw in equation (3.3.21), $\mathbb{E}[\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij}] = \sum_{i=1}^{n-1} \frac{1}{i}$. Using the fact that,

$$\sum_{i=1}^{n} \frac{1}{i} = \log n + \gamma + O(\frac{1}{n})$$

where $\gamma := \lim_{n\longrightarrow\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \log n\right)$ is the Euler constant, shows that $\left\{\mathbb{E}[\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij}] - \log n\right\}_{n\geq 1}$ is a convergent sequence. Now from (3.2.2), we have

$$\begin{aligned}
T_n^{NN} - \log n &\stackrel{d}{=} \sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij} - \mathbb{E}[\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij}] + \mathbb{E}[\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij}] - \log n + L_n^{\text{first}} + L_n^{\text{last}} \\
&\stackrel{d}{=} \sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij} - \mathbb{E}[\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij}] + \mathbb{E}[\sum_{i=1}^{n-1} \min_{i<j\leq n} L_{ij}] - \log n \\
&+ L_n^{\text{first}} + L_n^{\text{last}} - \mathbb{E}\left[L_n^{\text{first}}\right].
\end{aligned}$$

Therefore by using (3.3.22) and Proposition 3.2.1, we get $\left(T_n^{NN} - \log n\right)_{n\geq 1}$ converges weakly.

□

**Theorem 3.3.3.** *Assume that as* $t \longrightarrow 0+$, $\frac{f(t)}{t^\alpha} \longrightarrow C$, *where* $C > 0$ *is constant and* $0 < \alpha < 1$.

*Then as* $n \longrightarrow \infty$,

$$\frac{T_n^{NN}}{n^{1-\frac{1}{1+\alpha}}} \xrightarrow{\mathbb{P}} K_\alpha \tag{3.3.23}$$

*where*

$$K_\alpha := (\frac{1+\alpha}{C})^{\frac{1}{1+\alpha}} \frac{1+\alpha}{\alpha} \Gamma(1+\frac{1}{1+\alpha})$$

*and*

$$\mathbb{E}[T_n^{NN}] \sim K_\alpha n^{1-\frac{1}{1+\alpha}} \tag{3.3.24}$$

*Moreover, convergence in* (3.3.23) *happens in* $\mathcal{L}_2$.

*Proof.* Recall the double inequality (3.3.3) in the proof of Lemma 3.3.2. By the assumption of the theorem and (3.3.3), as $i \longrightarrow \infty$,

$$\frac{(\frac{C}{1+\alpha})^{\frac{1}{1+\alpha}} W_i}{(\frac{Y_i}{i})^{\frac{1}{1+\alpha}}} \longrightarrow 1 \quad a.s.$$

where $Y_i$'s are i.i.d. Exponential random variable each with mean one and $W_i = F^{-1}\left(1 - \exp(-\frac{Y_i}{i})\right)$. Therefore as $n \longrightarrow \infty$

$$\frac{(\frac{C}{1+\alpha})^{\frac{1}{1+\alpha}} \sum_{i=1}^{n-2} W_i}{\sum_{i=1}^{n-2} (\frac{Y_i}{i})^{\frac{1}{1+\alpha}}} \longrightarrow 1 \quad a.s.$$

Since $0 < \alpha < 1$ so $\frac{2}{1+\alpha} > 1$, thus $\mathsf{Var}\left(\sum_{i=1}^{n-2} (\frac{Y_i}{i})^{\frac{1}{1+\alpha}}\right)$ is uniformly bounded and so by the martingale convergence theorem $\sum_{i=1}^{n-2} (\frac{Y_i}{i})^{\frac{1}{1+\alpha}} - \mathbb{E}\left[\sum_{i=1}^{n-2} (\frac{Y_i}{i})^{\frac{1}{1+\alpha}}\right]$ converges almost surely. But

$$\mathbb{E}\left[\sum_{i=1}^{n-2} (\frac{Y_i}{i})^{\frac{1}{1+\alpha}}\right] = \Gamma(1+\frac{1}{1+\alpha}) \sum_{i=1}^{n-2} (\frac{1}{i})^{\frac{1}{1+\alpha}} .$$

Thus

$$\frac{\sum_{i=1}^{n-2} W_i}{K_\alpha n^{1-\frac{1}{1+\alpha}}} \longrightarrow 1 \quad a.s. \tag{3.3.25}$$

where

$$K_\alpha := (\frac{1+\alpha}{C})^{\frac{1}{1+\alpha}} \frac{1+\alpha}{\alpha}\Gamma(1+\frac{1}{1+\alpha})\,.$$

Now

$$\sum_{i=1}^{n-2} W_i - K_\alpha n^{1-\frac{1}{1+\alpha}} = \sum_{i=1}^{n-2} W_i - \mathbb{E}[\sum_{i=1}^{n-2} W_i] + \mathbb{E}[\sum_{i=1}^{n-2} W_i] - K_\alpha n^{1-\frac{1}{1+\alpha}}\,.$$

Recall that by Lemma 3.3.2, $\sum_{i=1}^{n-2} W_i - \mathbb{E}[\sum_{i=1}^{n-2} W_i]$ has an almost sure limit, so using (3.3.25) we get

$$\lim_{n\longrightarrow\infty} \frac{\mathbb{E}[\sum_{i=1}^{n-2} W_i]}{n^{1-\frac{1}{1+\alpha}}} = K_\alpha \tag{3.3.26}$$

and hence by Lemma 3.3.2, Lemma 3.3.1 and equation (3.3.1),

$$\mathbb{E}[T_n^{NN}] \sim K_\alpha n^{1-\frac{1}{1+\alpha}}.$$

Note that

$$\begin{aligned}
\mathbb{E}[\frac{T_n^{NN}}{n^{1-\frac{1}{1+\alpha}}} - K_\alpha]^2 &= \mathbb{E}[\frac{T_n^{NN} - \mathbb{E}[T_n^{NN}]}{n^{1-\frac{1}{1+\alpha}}} + \frac{\mathbb{E}[T_n^{NN}]}{n^{1-\frac{1}{1+\alpha}}} - K_\alpha]^2 \\
&= \frac{\mathbb{E}\left[\left(\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij} - \mathbb{E}[\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij}]\right)^2\right]}{(n^{1-\frac{1}{1+\alpha}})^2} \\
&+ \frac{\mathbb{E}\left[\left(L_n^{\text{first}} + L_n^{\text{last}} - \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}]\right)^2\right]}{(n^{1-\frac{1}{1+\alpha}})^2} + \left[\frac{\mathbb{E}[T_n^{NN}]}{n^{1-\frac{1}{1+\alpha}}} - K_\alpha\right]^2 \\
&= \frac{\mathsf{Var}\left[\sum_{i=2}^{n-1} \min_{i<j\le n} L_{ij}\right]}{(n^{1-\frac{1}{1+\alpha}})^2} + \frac{\mathsf{Var}\left[L_n^{\text{first}} + L_n^{\text{last}}\right]}{(n^{1-\frac{1}{1+\alpha}})^2}
\end{aligned}$$

$$+\left[\frac{\mathbb{E}[T_n^{NN}]}{n^{1-\frac{1}{1+\alpha}}}-K_\alpha\right]^2$$

converges to zero as $n \longrightarrow \infty$. Hence

$$\frac{T_n^{NN}}{n^{1-\frac{1}{1+\alpha}}} \xrightarrow{\mathbb{P}} K_\alpha$$

and in $\mathcal{L}_2$. $\square$

**Theorem 3.3.4.** *Let $-1 < \alpha < 0$ and assume that as $t \longrightarrow 0+$, $\frac{f(t)}{t^\alpha} \longrightarrow C$, where $C > 0$ is constant. Then the sequence $\{\mathbb{E}[T_n^{NN}]\}_{n\geq 1}$, is a convergent sequence and $T_n^{NN}$ converges weakly.*

*Proof.* As it has mentioned in the proof of Lemma 3.3.2, since $\frac{1}{1+\alpha} > 1$, we get

$$\sup_{n\geq 1} \mathsf{Var}(\sum_{i=1}^{n-2} W_i) < \infty\,.$$

Therefore $\sum_{i=1}^{n-2} W_i - \mathbb{E}[\sum_{i=1}^{n-2} W_i]$ as a martingale converges $a.s.$ and in $\mathcal{L}_2$. So by equation (3.3.1) and Proposition 3.2.1, $T_n^{NN} - \mathbb{E}[T_n^{NN}]$ converges weakly.

Now to complete the proof it is enough to show that $\left\{\mathbb{E}\left[T_n^{NN}\right]\right\}_{n\geq 1}$ is a convergent sequence. For that we apply Lemma 3.4.1 to get

$$\mathbb{E}[T_n^{NN}] = \int_0^\infty \frac{\left[\bar{F}(t)\right]^2\left[1-\left(\bar{F}(t)\right)^{n-2}\right]}{F(t)}dt + \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}]\,. \tag{3.3.27}$$

Now fix $\epsilon > 0$ and get $\delta > 0$ such that the equations leading to the double inequality (3.3.3) holds. Also find $M > 0$ such that $F(M) \geq \frac{1}{2}$. Consider the function $G : [0,\infty) \to [0,\infty)$ defined as

$$G(t) := \begin{cases} \frac{1}{F(t)} & \text{if } 0 < t < \delta \\ \frac{1}{F(\delta)} & \text{if } \delta \leq t \leq M \\ 2\bar{F}(t) & \text{otherwise} \end{cases}\,.$$

Then for any $n > 1$ and $t > 0$ we have

$$\frac{\left[\bar{F}(t)\right]^2 \left[1 - \left(\bar{F}(t)\right)^{n-2}\right]}{F(t)} \leq G(t)\,.$$

Also note that $\int_M^\infty G(t)\,dt \leq 2\int_0^\infty \bar{F}(t)\,dt < \infty$ as $F$ is positively supported and has finite first moment. Further by the choice of $\delta$ we get that on $(0, \delta)$ the density $f$ is strictly positive and $F$ is strictly increasing. So

$$\begin{aligned}
\int_0^\delta G(t)\,dt &= \int_0^\delta \frac{dt}{F(t)} \\
&= \int_0^{F(\delta)} \frac{dw}{w\,f\left(F^{-1}(w)\right)} \qquad \text{[substitute } w = F(t)\text{]} \\
&\leq \kappa \int_0^1 \frac{1}{w^{1+\frac{\alpha}{1+\alpha}}}\,dw < \infty\,,
\end{aligned}$$

where $\kappa > 0$ is some constant and the last but one inequality follows by using the double inequality (3.3.3) and the final inequality holds because $-1 < \alpha < 0$. Thus we get that

$$\int_0^\infty G(t)\,dt < \infty\,.$$

So by the dominated convergence theorem we conclude that

$$\lim_{n\to\infty} \int_0^\infty \frac{\left[\bar{F}(t)\right]^2 \left[1 - \left(\bar{F}(t)\right)^{n-2}\right]}{F(t)} dt$$

exists. This along with Proposition 3.2.1 proves that $\left\{\mathbb{E}\left[T_n^{NN}\right]\right\}_{n\geq 1}$ is convergent sequence, which completes the proof of the theorem.

□

*Remark* 3.3.1. Our results cover the cases where $|\alpha| < 1$. Note that the case $\alpha \leq -1$ cannot happen, since $f$ is a density function. For $\alpha \geq 1$ we do not have any general result except for the particular choice of $F$, namely when $F$ is Weibull distribution with shape parameter $(1 + \alpha)$ and scale parameter 1, we show in the following proposition that after proper scaling, the weak limit distribution of $T_n^{NN}$ is Normal.

**Proposition 3.3.1.** *Let $\alpha \geq 1$ and for $1 \leq i \leq n-1$, the intercity distances $\{L_{ij}\}_{i<j\leq n}$ in mean field TSP be i.i.d. Weibull distribution with shape parameter $(1+\alpha)$ and scale parameter 1, i.e.,*

$$f(t) = (1+\alpha)t^{\alpha}e^{-t^{(1+\alpha)}}\mathbf{1}\left(t>0\right) .$$

*Then as $n \longrightarrow \infty$, for $\alpha > 1$*

$$\frac{T_n^{NN} - \mathbb{E}[T_n^{NN}]}{n^{\frac{1}{2}-\frac{1}{1+\alpha}}} \xrightarrow{d} N(0,\ \frac{\alpha+1}{\alpha-1}\sigma^2(\alpha)) \tag{3.3.28}$$

*and for $\alpha = 1$,*

$$\frac{T_n^{NN} - \mathbb{E}[T_n^{NN}]}{\sqrt{\log n}} \xrightarrow{d} N(0, \sigma^2(\alpha)) \tag{3.3.29}$$

*where $\sigma^2(\alpha) = \Gamma(\frac{2}{1+\alpha}+1) - \Gamma^2(1+\frac{1}{1+\alpha})$.*

*Proof.* By assumption that $F$ is Weibull distribution with shape parameter $(1+\alpha)$ and scale parameter 1, we get

$$F(x) = 1 - e^{-x^{1+\alpha}}, \quad x \geq 0$$

Therefore $F^{-1}(t) = [-\log(1-t)]^{\frac{1}{1+\alpha}}$, where $0 < t < 1$. Hence,

$$\begin{aligned}\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij} &\stackrel{d}{=} \sum_{i=1}^{n-2} W_i \\ &= \sum_{i=1}^{n-2} [-\log(e^{-\frac{Y_i}{i}})]^{\frac{1}{1+\alpha}} \\ &= \sum_{i=1}^{n-2} (\frac{Y_i}{i})^{\frac{1}{1+\alpha}}\end{aligned}$$

where $Y_i$'s are i.i.d. Exponential random variable each with mean one. Note that

$$\mu(\alpha) := \mathbb{E}\left[Y_i^{\frac{1}{1+\alpha}}\right] = \Gamma(1+\frac{1}{1+\alpha})$$

and

$$\sigma^2(\alpha) := \mathsf{Var}\left[Y_i^{\frac{1}{1+\alpha}}\right] = \Gamma(\frac{2}{1+\alpha}+1) - \Gamma^2(1+\frac{1}{1+\alpha})\,.$$

Let

$$V_i(\alpha) := \frac{Y_i^{\frac{1}{1+\alpha}} - \mathbb{E}[Y_i^{\frac{1}{1+\alpha}}]}{\sigma(\alpha) i^{\frac{1}{1+\alpha}} \sqrt{\sum_{i=1}^{n-2} (\frac{1}{i})^{\frac{2}{1+\alpha}}}}$$

and $Z_n(\alpha) = \sum_{i=1}^{n-2} V_i(\alpha)$. Observe that $\mathbb{E}[V_i(\alpha)] = 0$ and $\sum_{i=1}^{n-2} \mathsf{Var}[V_i(\alpha)] = 1$. Choose $\delta > 0$ such that $\delta > \alpha - 1$. So for some $M > 0$,

$$\sum_{i=1}^{n-2} \mathbb{E}\left[|V_i(\alpha)|^{2+\delta}\right] \leq \frac{M}{\sigma(\alpha)^{2+\delta}} \frac{1}{[\sum_{i=1}^{n-2}(\frac{1}{i})^{\frac{2}{1+\alpha}}]^{\frac{2+\delta}{2}}} \sum_{i=1}^{n-2} (\frac{1}{i})^{\frac{2+\delta}{1+\alpha}}\,.$$

Since $\frac{2}{1+\alpha} \leq 1$ and $\frac{2+\delta}{1+\alpha} > 1$, we have

$$\lim_{n\to\infty} \sum_{i=1}^{n-2} \mathbb{E}\left[|V_i(\alpha)|^{2+\delta}\right] = 0\,.$$

Hence Lyapunov condition is satisfied for $\alpha \geq 1$ and so $Z_n(\alpha)$ converges in distribution to a standard Normal random variable, as $n$ goes to infinity. Now by equation (3.2.2) we have

$$\begin{aligned}\frac{T_n^{NN} - \mathbb{E}[T_n^{NN}]}{n^{\frac{1}{2}-\frac{1}{1+\alpha}}} &\stackrel{d}{=} \frac{\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{1+\alpha}} - \mathbb{E}[\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{1+\alpha}}]}{\{\mathsf{Var}[\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{1+\alpha}}]\}^{1/2}} \frac{\{\mathsf{Var}[\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{1+\alpha}}]\}^{1/2}}{n^{\frac{1}{2}-\frac{1}{1+\alpha}}} \\ &+ \frac{L_n^{\text{first}} + L_n^{\text{last}} - \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}]}{n^{\frac{1}{2}-\frac{1}{1+\alpha}}}\,,\end{aligned}$$

and thus the proof of proposition for $\alpha > 1$ is completed by Proposition 3.2.1. Note that when

$\alpha = 1$, by equation (3.2.2) we get

$$T_n^{NN} - \mathbb{E}[T_n^{NN}] \stackrel{d}{=} \frac{\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{2}} - \mathbb{E}[\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{2}}]}{\{\mathsf{Var}[\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{2}}]\}^{1/2}}\{\mathsf{Var}[\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{2}}]\}^{1/2}$$
$$+ L_n^{\text{first}} + L_n^{\text{last}} - \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}]\,.$$

But,

$$\mathsf{Var}[\sum_{i=1}^{n-2}(\frac{Y_i}{i})^{\frac{1}{2}}] = \sigma^2(1)\sum_{i=1}^{n-2}\frac{1}{i}$$

Therefore by Proposition 3.2.1 and the fact that $\sum_{i=1}^{n-2}\frac{1}{i} \sim \log n$ we get,

$$\frac{T_n^{NN} - \mathbb{E}[T_n^{NN}]}{\sqrt{\log n}} \xrightarrow{d} N(0, \sigma^2(1))$$

□

## 3.4 Technical result

The following lemma gives an expression for the mean of $T_n^{NN}$ in terms of the distribution function $F$. Under some further assumption on $F$ it also shows how the behavior of $\mathbb{E}\left[T_n^{NN}\right]$ depends on the behavior of the density $f$ of $F$ near zero.

**Lemma 3.4.1.** *Consider a mean field TSP with i.i.d. edge weights with distribution $F$ which is supported on $[0,\infty)$. Then*

$$\mathbb{E}[T_n^{NN}] = \int_0^\infty \frac{\left[\bar{F}(t)\right]^2\left[1 - \left(\bar{F}(t)\right)^{n-2}\right]}{F(t)}dt + \mathbb{E}[L_n^{\mathit{first}} + L_n^{\mathit{last}}]\,.$$

*Moreover if $F$ admits a continuous density $f$ which is strictly positive on the support $[0,\infty)$ then*

$$\mathbb{E}[T_n^{NN}] = \int_0^1 \frac{(1-w)^2(1-[1-w]^{n-2})}{w}\frac{1}{f(F^{-1}(w))}dw + \mathbb{E}[L_n^{\mathit{first}} + L_n^{\mathit{last}}]\,.$$

*Proof.* Let $\bar{F}(t) = 1 - F(t)$. From equation (3.2.2) we have

$$\mathbb{E}[T_n^{NN}] = \sum_{i=2}^{n-1} \mathbb{E}[\min_{i<j\leq n} L_{ij}] + \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}]$$

But,

$$\mathbb{E}[\min_{i<j\leq n} L_{ij}] = \int_0^\infty [\bar{F}(t)]^{n-i} dt,$$

and hence

$$\mathbb{E}[T_n^{NN}] = \int_0^\infty \frac{\left[\bar{F}(t)\right]^2 \left[1 - \left(\bar{F}(t)\right)^{n-2}\right]}{F(t)} dt + \mathbb{E}[L_n^{\text{first}} + L_n^{\text{last}}]\,,$$

which proves the first part of the lemma.

Now if we assume that $F$ admits a continuous density $f$ which is strictly positive on the support $[0, \infty)$ then the second expression follows by changing the variable $w = F\,(t)$ in the first. □

## 3.5 Discussion

We end this chapter with the following two subsections.

### 3.5.1 Assumptions on distribution function $F$

In our theorems, we assumed that the second moment of $F$ exists. This assumption is not needed. The following lemma says that if $F$ is a positively supported distribution with finite $\beta^{\text{th}}$-moment then for any $k > \frac{2}{\beta}$ we must have $\mathbb{E}\left[\left(\min_{1\leq i\leq k} Z_i\right)^2\right] < \infty$ where $Z_1, Z_2, \ldots$ are i.i.d. $F$.

**Lemma 3.5.1.** *Suppose $Z$ is a non-negative random variable such that for some $\beta > 0$, $\mathbb{E}[Z^\beta] < \infty$. Then for any $k > \frac{2}{\beta}$ we have*

$$\int_0^\infty t \left\{\mathbb{P}(Z > t)\right\}^k \, dt < \infty\,.$$

The proof of this lemma follows easily from Markov's inequality, so we omit it here. Now as before let random variable $W_i = F^{-1}\left(1 - \exp(-\frac{Y_i}{i})\right)$ where $Y_i$'s are Exponential with mean one. We have assumed that $F$ has finite first moment so then by taking $k = 3$ in Lemma 3.5.1 above we can conclude that $W_i$ has finite second moment for $i \geq 3$. Thus under the assumptions of Lemma 3.3.2 and following the proof of this lemma we can conclude that $\sum_{i=k}^{n-2}(W_i - \mathbb{E}[W_i])$ converges almost surely and in $\mathcal{L}_2$. Thus all the results stated in Section 3.3 hold except those on $\mathcal{L}_2$ convergence.

### 3.5.2 The relation of the objective function with lower records

Recall the equation (3.2.2)

$$T_n^{NN} \overset{d}{=} \sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij} + L_n^{\text{first}} + L_n^{\text{last}}$$

The study of the asymptotic behavior of $\sum_{i=2}^{n-1} \min_{i<j\leq n} L_{ij}$, can give more information about the behavior of the $T_n^{NN}$ for large $n$. One way to look at this summation, is through looking at the sum of lower records. Let $\{X_i\}_{i\geq 0}$ , be a sequence of independent and identically distributed random variables with continuous distribution function $F$ on $[0, \infty)$. The random variable $X_j$ is called a lower record, if $X_j \leq min\{X_1, ..., X_{j-1}\}$. By convention, $X_0$ is the first lower record. Define $K_0 \equiv 1$, and for $n \geq 1$, $K_n = min\left\{j > K_{n-1} : X_j < X_{K_{n-1}}\right\}$. Then $\left\{R_n^L := X_{K_n}, n \geq 0\right\}$ is called the sequence of "lower records" from $F$. Define the random variable $D_n := K_n - K_{n-1}$ to be the number of trials to get a new record and $N_n$, the number of lower records among $X_1, X_2, ..., X_n$. Then one can write

$$T_n^{NN} \overset{d}{=} \sum_{i=1}^{N_n} R_i^L D_i$$

In (Bose et al., 2003), there is necessary and sufficient condition for the partial sums of lower records to converge almost surely to a proper random variable. In fact they have proved that $\sum_{n=1}^{\infty} R_n^L < \infty$ a.e. if and only if $\int_0^1 x\frac{F(dx)}{F(x)} < \infty$. In our case, we have a weighted partial sum and that sum is up to a random variable $N_n$. Therefore, an answer to the question whether

$\sum_{i=1}^{N_n} R_i^L D_i$ converges or not, can lead us to know more about the behavior of $T_n^{NN}$.

# Chapter 4

# Random geometric graph with Cantor distributed vertices[1]

## 4.1 Introduction

### 4.1.1 Background and motivation

As we mentioned in Subsection 1.2.3, a *random geometric graph* consists of a set of vertices, distributed randomly over some metric space, in which two distinct such vertices are joined by an edge, if the distance between them is sufficiently small. More precisely, let $V_n$ be a set of $n$ points in $\mathbb{R}^d$, distributed independently according to some distribution $F$ on $\mathbb{R}^d$. Let $r$ be a fixed positive real number. Then, random geometric graph $\mathcal{G} = \mathcal{G}(V_n, r)$ is a graph with vertex set $V_n$ where two vertices $\mathbf{v} = (v_1, \ldots, v_d)$ and $\mathbf{u} = (u_1, \ldots, u_d)$ in $V_n$ are adjacent if and only if $\|\mathbf{v} - \mathbf{u}\| \leq r$ where $\|.\|$ is some norm on $\mathbb{R}^d$.

A considerable amount of work has been done on the *connectivity threshold* defined as

$$R_n = \inf \left\{ r > 0 \,\middle|\, \mathcal{G}(V_n, r) \text{ is connected} \right\} . \tag{4.1.1}$$

The case when the vertices are assumed to be uniformly distributed on $[0,1]^d$, Appel and Russo

[1]This chapter is based on the paper by Bandyopadhyay and Sajadi (2012b)

(2002) showed that with probability one

$$\lim_{n\to\infty} \frac{n}{\log n} R_n^d = \begin{cases} 1 & \text{for } d = 1, \\ \frac{1}{2d} & \text{for } d \geq 2 \end{cases}, \tag{4.1.2}$$

when the norm $\|.\|$ is taken to be the $\mathcal{L}_\infty$ or the sup norm. Later Penrose (2003) showed that the limit in (4.1.2) holds but with different constants for any $\mathcal{L}_p$ norm for $1 \leq p \leq \infty$. Penrose (1999) considered the case when the distribution $F$ has a continuous density $f$ with respect to the Lebesgue measure which remains bounded away from $0$ on the support of $F$. Under certain technical assumptions such as smooth boundary for the support, he showed that with probability one,

$$\lim_{n\to\infty} \frac{n}{\log n} R_n^d = C$$

where $C$ is an explicit constant which depends on the dimension $d$ and essential infimum of $f$ and its value on the boundary of the support. Recently, Sarkar and Saurabh (2010) [personal communication], studied a case when the density $f$ of the underlying distribution may have minimum zero. They in particular, proved that when the support of $f$ is $[0, 1]$ and $f$ is bounded below on any compact subset not containing origin but it is regularly varying at the origin, then

$$\frac{R_n}{F^{-1}(1/n)} \Longrightarrow Y_{1+\alpha}$$

where

$$Y_\alpha := \sup\left\{ S_{n+1}^{1/\alpha} - S_n^{1/\alpha} : n \geq 0 \right\}$$

and for $n \geq 1, S_n = \sum_{i=1}^{n} X_i$ where $X_i$'s are i.i.d. Exponential random variables with mean one and $S_0 = 0$. The proof by Sarkar and Saurabh (2010) can easily be generalized to the case where the density is zero at finitely many points. A question then naturally arises what happens to the case when the distribution function is flat on some intervals, that is, if density exists then it will be zero on some intervals. Also what happens in the some what extreme case, when the density may not exist even though the distribution function is continuous and has flat parts.

To consider these questions, in this chapter, we study the connectivity of random geometric graphs where the underlying distribution of the vertices has no mass and is also singular with respect to the Lebesgue measure, that is, it has no density. For that, we consider the *generalized Cantor distribution* with parameter $\phi$ denoted by $Cantor(\phi)$ as the underlying distribution of the vertices of the graph. The distribution function is then flat on infinitely many intervals. See Subsection 1.2.4 for the definition of Cantor distribution. We will show that the connectivity threshold converges almost surely to the length of the largest flat part of the distribution function and we also provide some finer asymptotic of the same.

Before we state the main results, we give a brief description of the Hausdorff dimension, based on (Falconer, 1986). Let $\{U_i\}$ be a $\delta$-cover of a set $U$, i.e., $U \subset \bigcup_i U_i$ and $0 < |U_i| \leq \delta$, $\forall i$, where for a non-empty subset $A$ of $\mathbb{R}^n$, $|A| = \sup\{|x-y| : x, y \in A\}$. Let $E \subset \mathbb{R}^n$ and let $d$ be a non-negative number. For $\delta > 0$ define

$$\mathcal{H}^d_\delta(E) = \inf \sum_{i=1}^{\infty} |U_i|^d,$$

where the infimum is over all (countable) $\delta$-covers $\{U_i\}$ of $E$. To get the *Hausdorff d-dimension* outer measure of $E$ (defined by $\mathcal{H}^d(E)$), we let $\delta \longrightarrow 0$. Thus, $\mathcal{H}^d(E) = \lim_{\delta \longrightarrow 0} \mathcal{H}^d_\delta(E)$. This limit exists, but maybe infinite. The restriction of $\mathcal{H}^d$ to the $\sigma-$field of $\mathcal{H}^d$-measurable sets is called Hausdorff d-dimension measure. There is a unique value, $d_E$, called the *Hausdorff dimension* of $E$ such that,

$$\mathcal{H}^d(E) = \infty \quad \text{if} \quad 0 \leq d < d_E$$

and

$$\mathcal{H}^d(E) = 0 \quad \text{if} \quad d_E < d < \infty.$$

Define $d_\phi$ to be the Hausdorff dimension of generalized Cantor set. It is known that for the standard Cantor set, this dimension is $\frac{\log 2}{\log 3}$ (see Theorem 2.1 of Chapter 7 of Stein and Shakarchi, 2005). Also, for generalized Cantor set, $d_\phi$ is given by $d_\phi = -\frac{\log 2}{\log \phi}$ (see Exercise 8 of Chapter 7 of Stein and Shakarchi, 2005). Note that the standard Cantor set is a special case when $\phi = 1/3$.

## 4.2 Main results

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with Cantor$(\phi)$ distribution on $[0,1]$. Given the graph $\mathcal{G} = \mathcal{G}(V_n, r)$, where $V_n = \{X_1, X_2, \ldots, X_n\}$, let $R_n$ be defined as in (4.1.1).

**Theorem 4.2.1.** *For any* $0 < \phi < 1/2$*, as* $n \longrightarrow \infty$ *we have*

$$R_n \longrightarrow 1 - 2\phi \text{ a.s.} \tag{4.2.1}$$

*Proof.* We draw a sample of size $n$ from $Cantor(\phi)$ on $[0,1]$. Let $N_n$ be the number of elements falling in the subinterval $[0,\phi]$ and $n - N_n$ in $[1-\phi, 1]$. From the construction $N_n \sim$ Bin $\left(n, \frac{1}{2}\right)$. In selecting this sample of size $n$, there are three cases which may happen. Some of these points may fall in interval $[0,\phi]$ and rest in interval $[1-\phi, 1]$. That means $N_n \notin \{0, n\}$. In this case the distance between the points in $[0,\phi]$ and $[1-\phi, 1]$ is at least $1 - 2\phi$. The other cases are when all points fall in $[0,\phi]$ or all fall in $[1-\phi, 1]$ , which in this case $N_n = n$ or $N_n = 0$. Let $m_n = \min\limits_{1 \leq i \leq n} X_i$, $M_n = \max\limits_{1 \leq i \leq n} X_i$ and we define

$$L_n := \max\left\{X_i \middle| \, 1 \leq i \leq n \text{ and } X_i \in [0,\phi]\right\} \tag{4.2.2}$$

and

$$U_n := \min\left\{X_i \middle| \, 1 \leq i \leq n \text{ and } X_i \in [1-\phi, 1]\right\} . \tag{4.2.3}$$

We will take $L_n = 0$ (and similarly $U_n = 0$) if the corresponding set is empty.

Now find a $K \equiv K(\phi)$ such that $\phi^K < \frac{1}{2}(1-\phi)(1-2\phi)$. Note that such a $K < \infty$ exists since $0 < \phi < 1$. Let $I_1, I_2, \ldots, I_{2^K}$ be the $2^K$ sub-intervals of length $\phi^K$ which are part of the $K^{\text{th}}$ stage of the "removal of middle interval" for obtaining the generalized Cantor set with parameter $\phi$. For $1 \leq j \leq 2^K$ define $N_j := \sum_{i=1}^n \mathbf{1}(X_i \in I_j)$, which is the number of sample points in the sub-interval $I_j$. From the construction of the the generalized Cantor distribution with parameter $\phi$ it follows that

$$\boldsymbol{N}_K := (N_1, N_2, \dots, N_{2^K}) \sim \text{Multinomial}\left(n; \left(\frac{1}{2^K}, \frac{1}{2^K}, \cdots, \frac{1}{2^K}\right)\right) , \tag{4.2.4}$$

and $N_n^{[0,\phi]} = \sum_{I_j \subseteq [0,\phi]} N_j$. Consider the event $E_n := \bigcap_{j=1}^{2^K} [N_j \geq 1]$. Observe that on the event $E_n$ the maximum inter point distance between two points in $[0, \phi]$ as well as in $[1-\phi, 1]$ is at most $2\phi^K + \phi(1-2\phi) < 1 - 2\phi$ by the choice of $K$. Thus on $E_n$ we must have $R_n = U_n - L_n$ and so we can write

$$R_n = (U_n - L_n)\ \mathbf{1}_{E_n} + R_n^* \,\mathbf{1}_{E_n^c} \tag{4.2.5}$$

where $R_n^*$ a is random variable such that $0 < R_n^* < \phi$ a.s. Observe that conditioned on $[N_1 = r_1, N_2 = r_2, \dots, N_{2^K} = r_{2^K}]$ we have $U_n \stackrel{d}{=} 1 - \phi + \phi m_{n-k}$ and $L_n \stackrel{d}{=} \phi M_k$ and $N_n^{[0,\phi]} = k$ where $k = \sum_{I_j \subseteq [0,\phi]} r_j$. More generally

$$\left((L_n, U_n), \boldsymbol{N}_K\right)_{n\geq 1} \stackrel{d}{=} \left(\left(\phi M_{N_n^{[0,\phi]}}, 1 - \phi + \phi m_{n-N_n^{[0,\phi]}}\right), \boldsymbol{N}_K\right)_{n\geq 1} . \tag{4.2.6}$$

Note that to be technically correct we define $M_0 = m_0 = 0$.

Now it is easy to see that $m_n \longrightarrow 0$ and $M_n \longrightarrow 1$ a.s. But by the SLLN, $N_n^{[0,\phi]}/n \longrightarrow 1/2$ a.s., thus both $\left(N_n^{[0,\phi]}\right)$ and $\left(n - N_n^{[0,\phi]}\right)$ are two subsequences which are converging to infinity a.s. Moreover

$$\mathbb{P}\left(E_n^c\right) \leq \sum_{j=1}^{2^K} \mathbb{P}\left(N_j = 0\right) = 2^K \left(1 - \frac{1}{2^K}\right)^n = 2^K \exp\left(-\alpha_K n\right) , \tag{4.2.7}$$

where $\alpha_K = -\log\left(1 - \frac{1}{2^K}\right) > 0$. Thus $\sum_{n=1}^{\infty} \mathbb{P}\left(E_n^c\right) < \infty$, so by the First Borel-Cantelli Lemma we have

$$\mathbb{P}\left(E_n^c \text{ infinitely often }\right) = 0 \;\Rightarrow\; \mathbb{P}\left(E_n \text{ eventually }\right) = 1 .$$

In other words $\mathbf{1}_{E_n} \longrightarrow 1$ a.s. and $\mathbf{1}_{E_n^c} \longrightarrow 0$ a.s. Finally observing that $0 \leq R_n^* \leq \phi$ we get from equations (4.2.5) and (4.2.6),

$$R_n \longrightarrow (1 - 2\phi) \ .$$

$\square$

Our next theorem gives finer asymptotic but before we state the theorem, we provide here some basic notations and facts. Let $m_n := \min\{X_1, X_2, \ldots, X_n\}$. Recall the equation (1.2.2) in the Subsection (1.2.4). Therefore we get

$$m_n \stackrel{d}{=} \begin{cases} \phi m_k & \text{with probability} \quad 2^{-n}\binom{n}{k} \text{ for } k = 1, 2, ..., n \\ \phi m_n + 1 - \phi & \text{with probability} \quad 2^{-n} \end{cases} \tag{4.2.8}$$

Let $a_n := \mathbb{E}[m_n]$. Using (4.2.8) Hosking (1994) derived the following recursion formula for the sequence $(a_n)$

$$(2^n - 2\phi)\, a_n = 1 - \phi + \phi \sum_{k=1}^{n-1} \binom{n}{k} a_k, \quad n \geq 1 \tag{4.2.9}$$

Moreover Knopfmacher and Prodinger (1996) showed that whenever $0 < \phi < 1/2$ then as $n \to \infty$,

$$\frac{a_n}{n^{-\frac{1}{d_\phi}}} \longrightarrow C\left(\phi\right) \ , \tag{4.2.10}$$

where

$$C\left(\phi\right) := \frac{(1-\phi)(1-2\phi)}{\phi \log 2} \Gamma(-\log_2 \phi) \zeta(-\log_2 \phi)\, , \tag{4.2.11}$$

and $d_\phi = -\frac{\log 2}{\log \phi}$ is the Hausdorff dimension of the generalized Cantor set. Here $\Gamma(\cdot)$ and $\zeta(\cdot)$ are the Gamma and Riemann zeta functions, respectively.

Our next theorem gives the rate convergence of $R_n$ to $(1 - 2\phi)$ in terms of the $\mathcal{L}_1$ norm.

**Theorem 4.2.2.** *For any* $0 < \phi < 1/2$*, as* $n \longrightarrow \infty$ *we have*

$$\frac{\left\| R_n - (1 - 2\phi) \right\|_1}{n^{-\frac{1}{d_\phi}}} \longrightarrow 2C\left(\phi\right) \ , \tag{4.2.12}$$

*where* $C\left(\phi\right)$ *is as in equation* (4.2.11) *and* $\left\|\cdot\right\|_1$ *is the* $\mathcal{L}_1$ *norm.*

*Proof.* Let $R_n^*$, $E_n$ and $N_n^{[0,\phi]}$ be as defined in the proof of the Theorem 4.2.1. Observe that

$$
\begin{aligned}
& \mathbb{E}\left[\left|R_n-(1-2\phi)\right|\right] \\
= \;& \mathbb{E}\left[\left(R_n-(1-2\phi)\right)\mathbf{1}_{E_n}\right]+\mathbb{E}\left[\left|R_n^*-(1-2\phi)\right|\mathbf{1}_{E_n^c}\right] \\
= \;& \mathbb{E}\left[\left(U_n-L_n-(1-2\phi)\right)\mathbf{1}_{E_n}\mathbf{1}_{2^K\le N_n^{[0,\phi]}\le n-2^K}\right]+\mathbb{E}\left[\left|R_n^*-(1-2\phi)\right|\mathbf{1}_{E_n^c}\right] \\
= \;& \mathbb{E}\left[\left(U_n-L_n-(1-2\phi)\right)\mathbf{1}_{2^K\le N_n^{[0,\phi]}\le n-2^K}\right] \\
& -\mathbb{E}\left[\left(U_n-L_n-(1-2\phi)\right)\mathbf{1}_{E_n^c}\mathbf{1}_{2^K\le N_n^{[0,\phi]}\le n-2^K}\right]+\mathbb{E}\left[\left|R_n^*-(1-2\phi)\right|\mathbf{1}_{E_n^c}\right] \\
= \;& \mathbb{E}\left[\left(U_n-L_n-(1-2\phi)\right)\mathbf{1}_{1\le N_n^{[0,\phi]}\le n-1}\right] \\
& -\mathbb{E}\left[\left(U_n-L_n-(1-2\phi)\right)\mathbf{1}_{E_n^c}\mathbf{1}_{1\le N_n^{[0,\phi]}\le n-1}\right]+\mathbb{E}\left[\left|R_n^*-(1-2\phi)\right|\mathbf{1}_{E_n^c}\right] .
\end{aligned}
\tag{4.2.13}
$$

In above the first equality holds because of (4.2.5) and the fact that on the event $E_n$ we must have $R_n > 1-2\phi$. Second, third and fourth equalities follows from the simple fact that $E_n \subseteq \left[2^K \le N_n^{[0,\phi]} \le n-2^K\right]$.

Now recall that $a_n = \mathbb{E}[m_n]$, so for the first part of the right-hand side of the equation (4.2.13) we can write

$$
\begin{aligned}
& \mathbb{E}\left[\left(U_n-L_n-(1-2\phi)\right)\mathbf{1}_{1\le N_n^{[0,\phi]}\le n-1}\right] \\
= \;& \frac{\phi}{2^n}\sum_{k=1}^{n-1}\binom{n}{k}\left(a_{n-k}+a_k\right) \\
= \;& \frac{1}{2^{n-1}}\left[\left(2^n-2\phi\right)a_n-(1-\phi)\right] ,
\end{aligned}
\tag{4.2.14}
$$

where the last equality follows from (4.2.9). The other two parts of the right-hand side of the equation (4.2.13) are bounded in absolute value by

$$
\mathbb{P}\left(E_n^c\right) \le 2^K \exp\left(-\alpha_K n\right)
$$

because of (4.2.7). Now observe that from equation (4.2.10) we get that $a_n \sim C(\phi)\, n^{-\frac{1}{d_\phi}}$ where $d_\phi = -\frac{\log 2}{\log \phi}$ is the Hausdorff dimension of the generalized Cantor set. Thus using (4.2.13) and

(4.2.14) we conclude that

$$\frac{\mathbb{E}\left[|R_n - (1-2\phi)|\right]}{a_n} \longrightarrow 2 \quad \text{as} \quad n \longrightarrow \infty\,.$$

This completes the proof using (4.2.10). □

## 4.3 Discussion

Consider the standard Cantor distribution. According to Theorem 4.2.2, the $\mathcal{L}_1$-norm of $\frac{R_n - 1/3}{a_n}$ converges to 2. The question now naturally arises is whether this convergence can also be in probability. For that, we need to check whether the ratio $\frac{\mathbb{E}[m_n^2]}{a_n^2}$ converges to 1. This is because as we have seen in the proof of Theorem 4.2.1 (see equations (4.2.5) and (4.2.6)), for $\phi = 1/3$,

$$R_n \stackrel{d}{=} \left[\frac{1}{3} + \frac{1}{3}\left(m_{n-N_n^{[0,\phi]}} + m_{N_n^{[0,\phi]}}\right)\right] \mathbf{1}_{E_n} + R_n^* \,\mathbf{1}_{E_n^c}\,. \tag{4.3.1}$$

Therefore,

$$\begin{aligned}\left(\frac{R_n}{a_n}\right) &\stackrel{d}{=} \left[\frac{1}{3a_n} + \frac{1}{3}\left(\frac{m_{n-N_n^{[0,\phi]}} + m_{N_n^{[0,\phi]}}}{a_n}\right)\right] \mathbf{1}_{E_n} + \frac{R_n^*}{a_n}\mathbf{1}_{E_n^c} \\ &\stackrel{d}{=} \left[\frac{1}{3a_n} + \frac{1}{3}\left(\frac{m_{n-N_n^{[0,\phi]}}}{a_{n-N_n^{[0,\phi]}}}\frac{a_{n-N_n^{[0,\phi]}}}{a_n} + \frac{m_{N_n^{[0,\phi]}}}{a_{N_n^{[0,\phi]}}}\frac{a_{N_n^{[0,\phi]}}}{a_n}\right)\right] \mathbf{1}_{E_n} + \frac{R_n^*}{a_n}\,\mathbf{1}_{E_n^c}\,.\end{aligned}$$

Now using the fact that $a_n \sim n^{-\log_2 3}$ and also $N_n^{[0,\phi]}/n \longrightarrow 1/2$ a.s. and $0 \le R_n^* \le \frac{1}{3}$ a.s., all we need to check is whether $\frac{m_n}{a_n}$ converges almost surely to 1 or not. Put $b_n := \mathbb{E}[m_n^2]$ and note that for $\epsilon > 0$,by Chebyshev's inequality

$$\mathbb{P}\left(|\frac{m_n}{a_n} - 1| > \epsilon\right) \le \frac{b_n - a_n^2}{\epsilon^2 a_n^2}\,.$$

Unfortunately, numerical result show that the ratio $\frac{b_n}{a_n^2}$ does not converge to 1. As $n$ increases, the ratio $\frac{b_n}{a_n^2}$ also increases. For example for $n = 1000, \phi = \frac{1}{3}$, the value of this ratio is 3.85. Table 4.1, presents values of this ratio for different values of $n$ and also $\phi$. Note that from the equation (4.2.9) we get

$$a_n = \frac{1-\phi}{2^n - 2\phi} + \frac{\phi}{2^n - 2\phi} \sum_{k=1}^{n-1} \binom{n}{k} a_k$$

and from the equation (4.2.8) we get ,

$$b_n = \frac{(1-\phi)^2}{2^n - 2\phi^2} + \frac{\phi^2}{2^n - 2\phi^2} \sum_{k=1}^{n-1} \binom{n}{k} b_k + \frac{2\phi(1-\phi)}{2^n - 2\phi^2} a_n \, .$$

| n | $\phi = \frac{1}{3}$ | $\phi = \frac{1}{4}$ | $\phi = \frac{1}{5}$ | $\phi = \frac{1}{6}$ | $\phi = \frac{1}{10}$ | $\phi = \frac{1}{100}$ | $\phi = \frac{1}{1000}$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.5 | 1.6 | 1.67 | 1.71 | 1.82 | 1.98 | 2 |
| 2 | 1.94 | 2.28 | 2.52 | 2.71 | 3.14 | 3.9 | 3.99 |
| 3 | 2.3 | 2.94 | 3.48 | 3.93 | 5.12 | 7.63 | 7.96 |
| 4 | 2.56 | 3.51 | 4.42 | 5.25 | 7.76 | 14.77 | 15.87 |
| 5 | 2.74 | 3.93 | 5.18 | 6.43 | 10.84 | 28.13 | 31.58 |
| 6 | 2.85 | 4.2 | 5.71 | 7.33 | 13.87 | 52.38 | 62.69 |
| 7 | 2.93 | 4.36 | 6.03 | 7.89 | 16.36 | 94.53 | 123.97 |
| 8 | 3 | 4.48 | 6.23 | 8.21 | 18.02 | 163.43 | 243.81 |
| 9 | 3.06 | 4.58 | 6.36 | 8.4 | 18.88 | 267.22 | 475.61 |
| 10 | 3.12 | 4.69 | 6.5 | 8.55 | 19.21 | 407.86 | 916.79 |
| 30 | 3.57 | 5.77 | 8.48 | 11.68 | 29.43 | 1370.7 | 28406.11 |
| 70 | 3.73 | 6.19 | 9.32 | 13.11 | 34.96 | 2510.13 | 188579.89 |
| 140 | 3.79 | 6.36 | 9.67 | 13.73 | 37.89 | 3236.76 | 403229.41 |
| 200 | 3.81 | 6.42 | 9.83 | 14.05 | 38.69 | 4566.98 | 274654.83 |
| 250 | 3.82 | 6.43 | 9.84 | 14.07 | 39.63 | 4117.59 | 452071.68 |
| 300 | 3.83 | 6.46 | 9.87 | 14.08 | 39.26 | 3890.95 | 540362.35 |
| 350 | 3.83 | 6.48 | 9.93 | 14.19 | 39.16 | 4474.16 | 462054.02 |
| 450 | 3.83 | 6.48 | 9.96 | 14.29 | 40.26 | 4950.65 | 392114.82 |
| 550 | 3.84 | 6.49 | 9.95 | 14.24 | 40.32 | 4292.15 | 590138.17 |
| 600 | 3.84 | 6.5 | 9.96 | 14.24 | 40.03 | 4202.98 | 625579.2 |
| 700 | 3.84 | 6.51 | 10.01 | 14.32 | 39.79 | 4668.34 | 558303.69 |
| 750 | 3.84 | 6.52 | 10.02 | 14.36 | 39.94 | 4979.78 | 499881.31 |
| 850 | 3.84 | 6.51 | 10.03 | 14.4 | 40.47 | 5265.81 | 425619.25 |
| 900 | 3.85 | 6.51 | 10.02 | 14.4 | 40.71 | 5209.37 | 432379.16 |
| 950 | 3.85 | 6.51 | 10.01 | 14.38 | 40.85 | 5061.99 | 467998.79 |
| 1000 | 3.85 | 6.51 | 10 | 14.36 | 40.9 | 4870.52 | 520766.74 |

Table 4.1: Values of $\frac{b_n}{a_n^2}$ for different $n$ and $\phi$

At the end of this section, it is worth noting that our proofs for Theorem 4.2.1 and Theorem 4.2.2, depend on the recursive nature of the generalized Cantor distribution (see equation (1.2.2)). Thus unfortunately, they do not have obvious extensions for other singular distributions. It will be interesting to derive a version of Theorem 4.2.1 for a general singular distribution with no mass and flat parts. Intuitively it seems that the final limit should be the length of the longest flat part. It will be more interesting if Theorem 4.2.2 can also be generalized for general singular distributions with no mass and flat parts where $(1 - 2\phi)$ is replaced by the length of the longest flat part and $d_\phi$ is replaced by the Hausdorff dimension of the support.

# Bibliography


D. Aldous and J. M. Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer, Berlin, 2004.

G. L. Alexanderson. About the cover: Euler and Königsberg's bridges: a historical view. *Bull. Amer. Math. Soc. (N.S.)*, 43(4):567–573, 2006.

M. J. B. Appel and R. P. Russo. The minimum vertex degree of a graph on uniform points in $[0, 1]^d$. *Adv. in Appl. Probab.*, 29(3):582–594, 1997.

M. J. B. Appel and R. P. Russo. The connectivity of a graph on uniform points on $[0, 1]^d$. *Statist. Probab. Lett.*, 60(4):351–357, 2002.

K. B. Athreya and P. E. Ney. *Branching processes*. Dover Publications Inc., Mineola, NY, 2004. Reprint of the 1972 original [Springer, New York; MR0373040].

D. Avis, A. Hertz, and O. Marcotte, editors. *Graph theory and combinatorial optimization*, volume 8 of *GERAD 25th Anniversary Series*. Springer, New York, 2005.

A. Bandyopadhyay and F. Sajadi. On the expected total number of infections for virus spread on a finite network. 2012a. URL `http://arxiv.org/abs/1202.5429`.

A. Bandyopadhyay and F. Sajadi. Connectivity threshold of random geometric graphs with Cantor distributed vertices. *Statist. Probab. Lett.*, 82:2103–2107, 2012b.

A. Bandyopadhyay and F. Sajadi. On the nearest neighbor algorithm for mean field traveling salesman problem(accepted for publication in journal of applied probability). 2013. URL `http://arxiv.org/abs/1206.6991`.

J. Bang-Jensen and G. Gutin. *Digraphs*. Springer Monographs in Mathematics. Springer-Verlag London Ltd., London, second edition, 2009. Theory, algorithms and applications.

A. L. Barabási, A. R. E. Crandall, and Reviewer. Linked: The new science of networks. *American Journal of Physics*, 71(4), 2003.

A. D. Barbour and S. Utev. Approximating the Reed-Frost epidemic process. *Stochastic Process. Appl.*, 113(2):173–197, 2004.

J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. *Proc. Cambridge Philos. Soc.*, 55:299–327, 1959.

M. Bellmore and G. L. Nemhauser. The traveling salesman problem: A survey. *Operations Res.*, 16:538–558, 1968.

B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.

A. Bose, S. Gangopadhyay, A. Sarkar, and A. Sengupta. Convergence of lower records and infinite divisibility. *J. Appl. Probab.*, 40(4):865–880, 2003.

G. Cantor. Über unendliche, lineare punktmannigfaltigkeiten v, [on infinite, linear point-manifolds (sets)]. *Math. Ann.*, 21:545–591, 1883.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, third edition, 2009.

M. Draief, A. Ganesh, and L. Massoulié. Thresholds for virus spread on networks. *Ann. Appl. Probab.*, 18(2):359–378, 2008.

P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

K. J. Falconer. *The geometry of fractal sets*, volume 85 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1986.

J. W. GavettBose. Three heuristic rules for sequencing jobs to a single production facility. *Management Science*, 11(8):166–176, 1965.

E. N. Gilbert. Random plane networks. *J. Soc. Indust. Appl. Math.*, 9:533–543, 1961.

G. Grimmett. *Percolation*, volume 321 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 1999.

P. Gupta and P. R. Kumar. Critical power for asymptotic connectivity in wireless networks. *In Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W. H. Fleming*, 1998.

W. Hamer. The milroy lectures on epidemic disease in englandthe evidence of variability and of persistency of type. *The Lancet*, 167(4306):655– 662, 1906. Originally published as Volume 1, Issue 4306.

J. R. M. Hosking. Moments of order statistics of the Cantor distribution. *Statist. Probab. Lett.*, 19(2):161–165, 1994.

S. Janson, T. Łuczak, and A. Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

D. S. Johnson and L. A. McGeoch. The traveling salesman problem: a case study. In *Local search in combinatorial optimization*, Wiley-Intersci. Ser. Discrete Math. Optim., pages 215–310. Wiley, Chichester, 1997.

W. O. Kermack and A. G. McKendrick. A contribution to mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Character*, 115(772), 1927.

A. Knopfmacher and H. Prodinger. Explicit and asymptotic formulae for the expected values of the order statistics of the Cantor distribution. *Statist. Probab. Lett.*, 27(2):189–194, 1996.

F. R. Lad and W. F. C. Taylor. The moments of the Cantor distribution. *Statist. Probab. Lett.*, 13 (4):307–310, 1992.

C. Lefèvre and S. Utev. Poisson approximation for the final state of a generalized epidemic process. *Ann. Probab.*, 23(3):1139–1162, 1995.

M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99:2566–2572, 2002.

C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Dover Publications Inc., Mineola, NY, 1998. ISBN 0-486-40258-4. Corrected reprint of the 1982 original.

M. D. Penrose. The longest edge of the random minimal spanning tree. *The Annals of Applied Probability*, 7, 1997.

M. D. Penrose. A strong law for the largest nearest-neighbour link between random points. *J. London Math. Soc. (2)*, 60(3):951–960, 1999.

M. D. Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.

D. J. Rosenkrantz, R. E. Stearns, and P. M. Lewis, II. An analysis of several heuristics for the traveling salesman problem. *SIAM J. Comput.*, 6(3):563–581, 1977.

A. Sarkar and B. Saurabh. Random geometric graphs with densities having a zero. *Unpublished results*, 2010.

H. J. S. Smith. On the integration of discontinuous functions. *Proc. Lond. Math. Soc.*, 6:140–153, 1875.

E. M. Stein and R. Shakarchi. *Real analysis*. Princeton Lectures in Analysis, III. Princeton University Press, Princeton, NJ, 2005. Measure theory, integration, and Hilbert spaces.

J. Wästlund. The mean field traveling salesman and related problems. *Acta Math.*, 204(1): 91–150, 2010.

J. Wästlund. Replica symmetry of the minimum matching. *Annals of Mathematics*, 175(3): 1061–1091, 2012.

D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(4): 440–442, June 1998.

N. Weaver, V. Paxson, S. Staniford, and R. Cunningham. A taxonomy of computer worms. In *Proceedings of the 2003 ACM Workshop on Rapid Malcode, WORM 2003, Washington, DC, USA, October 27*, 2003.

J. G. Wendel. Note on the gamma function. *Amer. Math. Monthly*, 55:563–564, 1948.