

# A Probabilistic Investigation of $k$ Nearest Neighbor Directed Graphs

by

**Selim Bahadır**

A Dissertation Submitted to the  
Graduate School of Sciences and Engineering  
in Partial Fulfillment of the Requirements for  
the Degree of

Doctor of Philosophy

in

Mathematics



July 20, 2016

# A Probabilistic Investigation of $k$ Nearest Neighbor Directed Graphs

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

**Selim Bahadır**

and have found that it is complete and satisfactory in all respects,  
and that any and all revisions required by the final  
examining committee have been made.

Committee Members:

---

Assoc. Prof. Elvan Ceyhan

---

Assoc. Prof. Mine Çağlar

---

Asst. Prof. Serdar Altok

---

Assoc. Prof. Atilla Yılmaz

---

Asst. Prof. Özgür Asar

Date: \_\_\_\_\_

## ABSTRACT

This thesis is devoted to the investigation of the asymptotic behaviour of quantities based on random  $k$  nearest neighbor ( $k$ NN) digraphs. The  $k$ NN digraph of a finite point set in  $\mathbb{R}^d$  is obtained by inserting arcs from each point to its  $k$  nearest neighbors (i.e.,  $k$  closest members in the set). We first study the number of copies of a given fixed digraph in  $k$ NN digraph of the data from a random point process in  $\mathbb{R}^d$ . Based on the asymptotic theory for functionals of point sets under homogeneous Poisson process and uniform binomial process, we provide a general result for the asymptotic behavior of the subdigraph counts; and as corollaries, we obtain asymptotic results for the number of vertices with fixed indegree, the number of shared  $k$ NN pairs and the number of reflexive  $k$ NNs in a  $k$ NN digraph. Under some special settings, we also derive the exact values of the means and variances of these quantities and examine stochastic dependence between them.  $k$ NN type relations are widely used in spatial data analysis e.g. for testing spatial patterns of segregation and association. In a data set consisting of multiple classes, spatial segregation occurs when members of a class tend to be found near members of the same class while spatial association occurs when members of a class tend to be found near members of the other class or classes. We study these patterns using  $k$ NN digraph of the data points. We construct a contingency table called  $k$  nearest neighbor contingency table ( $k$ NNCT) based on classifying the arcs with respect to the classes of their endpoints. We analyze asymptotic distribution of the cell counts in a  $k$ NNCT under four different random settings. For two of these random settings, data points are obtained from a marked point process whereas in the other two, points are assumed to be fixed but their class identities (or labels) are randomly assigned. For each underlying process, we derive exact or asymptotic values of the expectations, variances and covariances of the cell counts of the  $k$ NNCT and obtain asymptotic distribution of the cell counts and quadratic forms based on them. Moreover, we demonstrate that random  $k$ NN digraphs are isomorphism-invariant. We extend the classification of isomorphism-invariant random graphs to digraphs and along this line, we introduce new families of random digraphs according to where randomness resides in the structure, namely, arc random digraphs, vertex random digraphs and vertex-arc random digraphs. We introduce randomness in the direction of the edges of a random graph and obtain direction random digraphs as well. We study relations between these four random digraph families and also determine where random  $k$ NN digraphs fall according to this classification.

## ÖZETÇE

Bu tez, rastgele  $k$  en yakın komşu ( $k$ EYK) yönlü çizgelerini temel alan niceliklerin asimptotik davranışlarının araştırılması üzerinedir.  $\mathbb{R}^d$ 'deki bir sonlu noktalar kümesinin  $k$ EYK yönlü çizgesi, her noktadan en yakın  $k$  komşusuna (yani, kümedeki ona en yakın  $k$  noktaya) oklar çizilmesiyle elde edilir. İlk olarak, verilmiş sabit bir yönlü çizgenin bir rastgele nokta sürecinden elde edilmiş  $k$ EYK yönlü çizgedeki kopyalarının sayısını çalışıyoruz. Homojen Poisson süreci ve düzgün nokta süreci sonucu oluşan nokta kümeleri üzerine tanımlı fonksiyonellerle ilgili asimptotik teoriyi kullanarak, alt yönlü çizge sayılarının asimptotik davranışları üzerine genel sonuçlar sağlıyoruz. Bu sonuçlardan çıkarım olarak, bir  $k$ EYK yönlü çizgesindeki sabit iç dereceye sahip noktaların sayısı, müşterek  $k$ EYK ikililerinin sayısı ve karşılıklı  $k$ EYK ikililerinin sayısı için asimptotik sonuçlar elde ediyoruz. Bazı özel koşullar altında, bu sayımların ortalama ve varyans değerlerini hesaplıyoruz ve aralarındaki stokastik bağımlılıkları inceliyoruz.  $k$ EYK tipi bağıntılar uzaysal veri analizinde yaygın olarak kullanılmaktadır (örneğin, ayrışma ve ortaşım motiflerini test etmek için). Birden fazla sınıftan oluşan bir veri kümesinde, bir sınıfın elemanları aynı sınıftaki elemanların yakınında bulunuyorsa uzaysal ayrışma; eğer bir sınıfın elemanları başka sınıflardaki elemanların yakınında ise uzaysal ortaşım meydana gelir. Bu tarz motifleri, veri noktalarının  $k$ EYK yönlü çizgesini kullanarak çalışıyoruz. Okları, iki uç noktalarının ait oldukları sınıflara göre kategorize edip,  $k$  en yakın komşu çapraz tablo ( $k$ EYKÇT) adında bir çapraz çizelge oluturuyoruz. Bir  $k$ EYKÇT'deki girdilerin asimptotik dağılımını dört farklı rastgele model altında inceliyoruz. Bu dört rastgele modelin ikisinde veri noktaları bir etiketli nokta sürecinden elde edilirken diğer ikisinde ise, noktalar sabit kabul ediliyor ama noktaların sınıf etiketleri rastgele atanıyor. Bazı alınan her bir model için,  $k$ EYKÇT'deki girdilerin beklenti, varyans ve kovaryans değerlerini tam veya asimptotik olarak hesaplıyoruz ve tablo girdilerinin ve bunlara bağlı karesel ifadelerin asimptotik dağılımlarını elde ediyoruz. Ayrıca, rastgele  $k$ EYK yönlü çizgelerinin eşdönüşüm-değişmez olduğunu gösteriyoruz ve eşdönüşüm-değişmez rastgele çizgelerin sınıflandırılmasını yönlü çizgelere uyarlıyoruz. Bu doğrultuda, rastgeleliğin yapının neresinde yer aldığına göre yeni rastgele yönlü çizge kümeleri tanımlıyoruz; şöyle ki, ok rastgele yönlü çizgeleri, köşe rastgele yönlü çizgeleri ve ok-köşe rastgele yönlü çizgeleri. İlaveten, bir rastgele çizgenin kenarlarına rastgele yön atayarak yön rastgele yönlü çizgelerini tanımlıyoruz. Bu dört rastgele yönlü çizge kümesinin arasındaki ilişkileri çalışıyoruz ve ek olarak, rastgele  $k$ EYK yönlü çizgelerinin bu sınıflandırmada nereye düştüğünü belirliyoruz.

## ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Assoc. Prof. Elvan Ceyhan, for his guidance on the subject. I also thank my family for believing in me and supporting me with their love during my writing of this thesis. Finally, thanks to my fellow graduate students for making life more livable and enjoyable.



# TABLE OF CONTENTS

|                                                                                                                         |             |
|-------------------------------------------------------------------------------------------------------------------------|-------------|
| <b>Abstract</b>                                                                                                         | <b>iii</b>  |
| <b>Özetçe</b>                                                                                                           | <b>iv</b>   |
| <b>Acknowledgements</b>                                                                                                 | <b>v</b>    |
| <b>Table of Contents</b>                                                                                                | <b>viii</b> |
| <b>List of Tables</b>                                                                                                   | <b>ix</b>   |
| <b>List of Figures</b>                                                                                                  | <b>x</b>    |
| <b>List of Abbreviations</b>                                                                                            | <b>xi</b>   |
| <b>Chapter 1: Introduction</b>                                                                                          | <b>1</b>    |
| 1.1 Preliminaries . . . . .                                                                                             | 6           |
| 1.1.1 Graph Theoretical Tools . . . . .                                                                                 | 6           |
| 1.1.2 Probabilistic Tools . . . . .                                                                                     | 7           |
| <b>Chapter 2: Subdigraph Counts in Random <math>k</math>NN Digraphs</b>                                                 | <b>10</b>   |
| 2.1 Introduction . . . . .                                                                                              | 10          |
| 2.2 Proof of the Main Theorem . . . . .                                                                                 | 11          |
| 2.2.1 CLT Results . . . . .                                                                                             | 12          |
| 2.2.2 Convergence of Means and Complete Convergence Results . . . . .                                                   | 16          |
| 2.3 Number of Vertices with Indegree $j$ . . . . .                                                                      | 18          |
| 2.3.1 Upper Bound on the Indegree . . . . .                                                                             | 18          |
| 2.3.2 Asymptotic Distribution of $Q_j^{[k]}$ . . . . .                                                                  | 21          |
| 2.3.3 Joint Distribution of $Q_j^{[k]}$ 's . . . . .                                                                    | 24          |
| 2.4 Number of Shared $k$ NNs . . . . .                                                                                  | 26          |
| 2.5 Number of Reflexive $k$ NNs . . . . .                                                                               | 28          |
| 2.6 The Case of $d = k = 1$ . . . . .                                                                                   | 30          |
| 2.6.1 Means and Variances of $R^{[1]}$ and $Q^{[1]}$ . . . . .                                                          | 32          |
| 2.6.2 A Recurrence Relation for the Exact Distribution of $R^{[1]}(\mathcal{U}_n)$ . . . . .                            | 33          |
| 2.7 Pairwise Dependence of $R^{[1]}(\mathcal{U}_n)$ , $Q^{[1]}(\mathcal{U}_n)$ and $Q_j^{[1]}(\mathcal{U}_n)$ . . . . . | 34          |

|                    |                                                                                           |           |
|--------------------|-------------------------------------------------------------------------------------------|-----------|
| <b>Chapter 3:</b>  | <b><math>k</math> Nearest Neighbor Contingency Tables</b>                                 | <b>40</b> |
| 3.1                | Introduction . . . . .                                                                    | 40        |
| 3.2                | Marked Homogeneous Poisson Process . . . . .                                              | 41        |
| 3.2.1              | Means, Variances and Covariances of the Cell Counts under $(\mathcal{P}_n, \mathcal{M})$  | 41        |
| 3.2.2              | Asymptotic Distribution of the Cell Counts under $(\mathcal{P}_n, \mathcal{M})$ . . . .   | 43        |
| 3.3                | Marked Uniform Binomial Process . . . . .                                                 | 46        |
| 3.4                | Random Labelling . . . . .                                                                | 47        |
| 3.4.1              | Type 1 RL . . . . .                                                                       | 48        |
| 3.4.1.1            | Means, Variances and Covariances of the Cell Counts<br>under Type 1 RL . . . . .          | 48        |
| 3.4.1.2            | Asymptotic Distribution of the Cell Counts under Type<br>1 RL . . . . .                   | 49        |
| 3.4.2              | Type 2 RL . . . . .                                                                       | 52        |
| 3.4.2.1            | Means, Variances and Covariances of the Cell Counts<br>under Type 2 RL . . . . .          | 52        |
| 3.4.2.2            | Asymptotic Distributions of Diagonal Cell Counts under<br>Type 2 RL . . . . .             | 55        |
| 3.4.2.3            | Off-Diagonal Cell Counts . . . . .                                                        | 60        |
| <b>Chapter 4:</b>  | <b>Classification of Isomorphism-Invariant Random Digraphs</b>                            | <b>64</b> |
| 4.1                | Introduction . . . . .                                                                    | 64        |
| 4.2                | ARDs, VRDs and VARDs . . . . .                                                            | 66        |
| 4.2.1              | Arc Random Digraphs . . . . .                                                             | 66        |
| 4.2.2              | Vertex Random Digraphs . . . . .                                                          | 68        |
| 4.2.3              | Vertex-Arc Random Digraphs . . . . .                                                      | 69        |
| 4.3                | Inclusion/Exclusion Relations between ARDs, VRDs and VARDs . . . .                        | 70        |
| 4.4                | Direction Random Digraphs . . . . .                                                       | 74        |
| 4.4.1              | DERDs, DVRDs and DVERDs . . . . .                                                         | 75        |
| 4.5                | Inclusion/Exclusion Relations of DERDs with respect to VARDs . . . .                      | 76        |
| 4.6                | Where Do Random $k$ NN Digraphs Reside? . . . . .                                         | 83        |
| <b>Chapter 5:</b>  | <b>Conclusions</b>                                                                        | <b>87</b> |
| <b>Appendix A:</b> | <b>Proofs of Auxiliary Results for the Main Theorem</b>                                   | <b>90</b> |
| A.1                | Proof of Proposition 2.2.1 . . . . .                                                      | 90        |
| A.2                | Proof of Proposition 2.2.5 . . . . .                                                      | 91        |
| <b>Appendix B:</b> | <b>Derivation of Means and Variances of <math>R^{[1]}</math> and <math>Q^{[1]}</math></b> | <b>93</b> |
| B.1                | Proof of Theorem 2.6.2 . . . . .                                                          | 93        |

|                                                                          |                                                                           |            |
|--------------------------------------------------------------------------|---------------------------------------------------------------------------|------------|
| B.2                                                                      | Proof of Corollary 2.6.3 . . . . .                                        | 95         |
| B.3                                                                      | Proof of Theorem 2.6.4 . . . . .                                          | 95         |
| <b>Appendix C: Derivation of Means, Variances and Covariances of the</b> |                                                                           |            |
|                                                                          | <b>Cell Counts in <math>k</math>NNCT</b>                                  | <b>99</b>  |
| C.1                                                                      | Under $(\mathcal{P}_n, \mathcal{M})$ . . . . .                            | 99         |
| C.2                                                                      | Under Type 2 RL . . . . .                                                 | 107        |
| C.3                                                                      | Derivation of $\rho(\alpha\beta, \theta\gamma)$ under Type 2 RL . . . . . | 110        |
| <b>Bibliography</b>                                                      |                                                                           | <b>112</b> |



## LIST OF TABLES

|     |                                                                    |    |
|-----|--------------------------------------------------------------------|----|
| 2.1 | Values of $q(d, k)$ for $d = 1, 2$ and $k = 1, \dots, 5$ . . . . . | 27 |
| 2.2 | Values of $r(d, k)$ for $d = 1, 2$ and $k = 1, \dots, 5$ . . . . . | 29 |
| 3.1 | $k$ NNCT for $m$ classes. . . . .                                  | 41 |

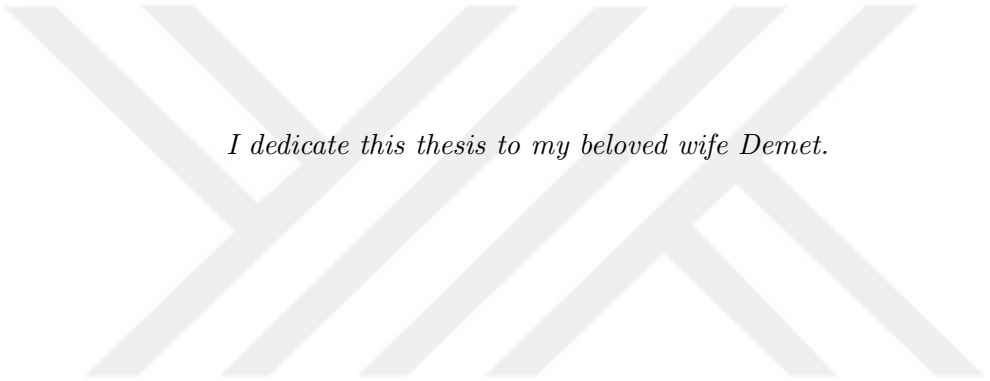


## LIST OF FIGURES

|     |                                                                                                                                                                                                                                                                                           |    |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | For $V = \{v_1, \dots, v_{10}\}$ , $D_{nn}^{[1]}(V)$ (on the left) and $D_{nn}^{[2]}(V)$ (on the right). . . . .                                                                                                                                                                          | 2  |
| 1.2 | Segregation of blue points and association of yellow and red points. . . . .                                                                                                                                                                                                              | 4  |
| 1.3 | 2NNCT of $V = V_1 \cup V_2 \cup V_3$ with $V_1 = \{v_1, v_2, v_3\}$ , $V_2 = \{v_4, v_5, v_6\}$ and $V_3 = \{v_7, v_8, v_9, v_{10}\}$ where $v_1, \dots, v_{10}$ are given in Figure 1.1. . . . .                                                                                         | 5  |
| 2.1 | An illustration of $n$ points with $d_{in}(v_0) = j$ for $d = 2$ , $n = 20$ , $k = 3$ and $j = 11$ . Notice that $s_1 = s_2 = s_3 = 3$ , $s_4 = 2$ and $s_5 = 0$ . . . . .                                                                                                                | 20 |
| 2.2 | An illustration of the event $E_1 \cap E_2$ for $d = 2$ and $k = 2$ . . . . .                                                                                                                                                                                                             | 21 |
| 2.3 | An illustration of the event $E_3$ for $d = 2$ , $k = 3$ , $j = 11$ , $s_1 = s_2 = s_3 = 3$ , $s_4 = 2$ and $s_5 = 0$ . . . . .                                                                                                                                                           | 23 |
| 2.4 | An illustration of the event $E_5$ for $d = 2$ and $k = 6$ . . . . .                                                                                                                                                                                                                      | 24 |
| 2.5 | An illustration of $n$ points with (a) $n/2$ reflexive pairs for $n = 22$ , (b) $(n - 1)/2$ reflexive pairs for $n = 23$ , and (c) $(n - 3)/2$ reflexive pairs for $n = 23$ . . . . .                                                                                                     | 35 |
| 2.6 | An illustration for $d = 2$ , $n = 18$ , $j = 4$ , $t_4 = 3$ and $s_4 = 3$ . Note that the sample points in the dashed circles are of indegree 4 and hence $Q_4 \geq 3$ . . . . .                                                                                                         | 37 |
| 3.1 | An illustration of $E_3$ for $d = 2$ and $k = 2$ . . . . .                                                                                                                                                                                                                                | 44 |
| 4.1 | Venn diagram of vertex-arc random digraphs for $n \geq 4$ . The results of this paper show that $\text{ARD} \cap \text{VRD} = \emptyset$ and all the four regions in the figure are nonempty for $n \geq 4$ . . . . .                                                                     | 74 |
| 4.2 | Venn diagram of direction random digraphs. The results of the paper imply that all the six regions in the figure are nonempty. In particular, the region $\text{DERD} \cap \text{DVRD}$ only consists of DRDs with $n \leq 3$ and DRDs generated by $\mathbf{G}(n, p_e = 1)$ . . . . .    | 77 |
| 4.3 | Venn diagram of DERDs and VARDs for $n \geq 4$ . The results in this paper indicate that all the five regions in the figure are nonempty. In addition, the intersection of the classes DERDs and VARDs is the family of ARDs, i.e., $\text{DERD} \cap \text{VARD} = \text{ARD}$ . . . . . | 81 |

## LIST OF ABBREVIATIONS

|        |                                      |
|--------|--------------------------------------|
| ARD:   | Arc Random Digraph                   |
| DERD:  | Direction-Edge Random Digraph        |
| DRD:   | Direction Random Digraph             |
| DVERD: | Direction-Vertex-Edge Random Digraph |
| DVRD:  | Direction-Vertex Random Digraph      |
| ERG:   | Edge Random Graph                    |
| GARD:  | Generalized Arc Random Digraph       |
| HPP:   | Homogeneous Poisson Process          |
| NN:    | Nearest Neighbor                     |
| NNCT:  | Nearest Neighbor Contingency Table   |
| RL:    | Random Labelling                     |
| RNND:  | Random Nearest Neighbor Digraph      |
| UBP:   | Uniform Binomial Process             |
| VARD:  | Vertex-Arc Random Digraph            |
| VERG:  | Vertex-Edge Random Graph             |
| VRD:   | Vertex Random Digraph                |
| VRG:   | Vertex Random Graph                  |



*I dedicate this thesis to my beloved wife Demet.*

## Chapter 1

## INTRODUCTION

The nearest neighbor (NN) relations and their properties have been extensively studied in various fields, such as probability and statistics (Friedman and Rafsky (1983), Bickel and Breiman (1983) and Avram and Bertsimas (1993)), computer science (Eppstein et al. (1997)), and ecology (Clark and Evans (1954), Cox (1981) and Pielou (1961)). Based on the NN type relations, graphs and digraphs are constructed and related graph quantities/invariants are widely studied (Cuzick and Edwards (1990), Eppstein et al. (1997), Penrose and Yukich (2001), Kozakova et al. (2006) and Wade (2007)). We consider  $k$ NN digraphs and study probabilistic aspects of quantities based on these digraphs under various random processes.

A *directed graph* (or simply *digraph*)  $D$  consists of a non-empty finite set  $V(D)$  of elements called *vertices* and a finite set  $A(D)$  of ordered pairs of distinct vertices called *arcs* (or *directed edges*). We call  $V(D)$  the vertex set and  $A(D)$  the arc set of  $D$ .

Let  $k, d \geq 1$  be fixed integers,  $n \geq k + 1$  be an integer and  $V = \{v_1, \dots, v_n\}$  be a finite set of points in  $\mathbb{R}^d$ . For any  $v \in V$ , let  $kNN(v)$  denote the set of  $k$  closest points to  $v$  among the points in  $V \setminus \{v\}$  with respect to a given metric and each point in  $kNN(v)$  is called a  $k$  nearest neighbor ( $k$ NN) of  $v$ . Throughout this thesis, we consider the usual Euclidean metric and denote the distance between the points  $u$  and  $v$  in  $\mathbb{R}^d$  as  $\|u - v\|$ . Obviously,  $kNN(v)$  may not be well defined if there exist ties in the  $k$ NN structure, i.e., there exist points  $u, w \in V$  such that  $\|u - v\| = \|w - v\|$ . However, such a tie occurs with probability zero for random point sets which are obtained from a process with almost surely continuous density, in particular from homogeneous Poisson process (HPP) or uniform binomial process (UBP). Hence we may always assume that pairwise distances are distinct and  $kNN(v)$  is well defined for the random point sets under consideration.

The  $k$ NN digraph on the point set  $V$ , denoted as  $D_{nn}^{[k]}(V)$ , is obtained by including an arc with tail  $v_i$  and head  $v_j$  whenever  $v_j$  is one of the  $k$ NNs of  $v_i$ . In other words,  $D_{nn}^{[k]}(V)$  is actually the digraph with vertex set  $V$  and arc set  $A = \{(v, u) : u, v \in V, u \in kNN(v)\}$ . See Figure 1.1 for 1NN and 2NN digraphs of a set of ten points.

The  $k$ NN digraph is also referred to as the  $k$ NN graph in literature (see, Eppstein et al. (1997)); but, since the  $k$ NN relation is not symmetric, we opt to use “ $k$ NN digraph” which reflects this asymmetry. Also, the underlying graph of a  $k$ NN digraph (an underlying graph of a digraph is obtained by replacing each arc with an (undirected)

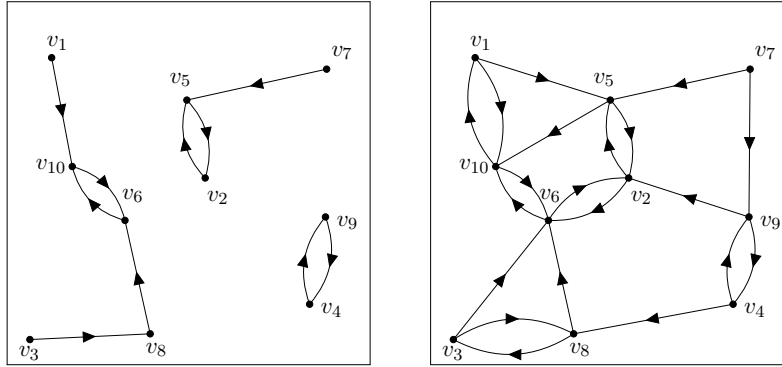


Figure 1.1: For  $V = \{v_1, \dots, v_{10}\}$ ,  $D_{nn}^{[1]}(V)$  (on the left) and  $D_{nn}^{[2]}(V)$  (on the right).

edge, disallowing multiple edges between two vertices (Chartrand and Lesniak (1996))) is referred to as the  $k$ NN graph (see, e.g., Friedman and Rafsky (1983), Penrose and Yukich (2001)). Throughout this thesis, we focus only on  $k$ NN digraphs.

For a fixed region in the space, we study the quantities based on the  $k$ NN digraphs of random point sets obtained from HPP of high intensity and large independent samples of non-random size from the uniform distribution on the region. Thus, we basically investigate random  $k$ NN digraphs and quantities based on them. Random graph models such as Erdős-Rényi graphs, random geometric graphs and nearest neighbor type graphs are used in various fields. The most frequently studied is the one proposed in Gilbert (1959), denoted  $\mathbf{G}(n, p)$ , in which each possible edge between  $n$  vertices occurs independently with probability  $p \in (0, 1)$ . However, in literature,  $\mathbf{G}(n, p)$  is usually called Erdős-Rényi model as they developed the theory (Erdős and Rényi (1960)). Another commonly considered model is random geometric graphs which are constructed by randomly placing a fixed number of vertices in some metric space (according to a specified probability distribution) and connecting two vertices by an edge if and only if their distance is smaller than a certain neighborhood radius. For more information about random geometric graphs, see Penrose (2003). The number of copies of a fixed graph in random graphs are analyzed by many authors (e.g., Nowicki and Wierman (1988), Ruciński (1988) and Janson et al. (2004) for  $\mathbf{G}(n, p)$  and Najim and Russo (2003), Yu (2009) and Shang (2010) for random geometric graphs).

We mainly focus on asymptotic distribution of the number of copies of a fixed digraph in random  $k$ NN digraphs. A digraph is called *weakly connected* if replacing all of its arcs with undirected edges produces a connected graph. By using the results of Penrose and Yukich (2001, 2002), we obtain asymptotic results for the number of subdigraphs of random  $k$ NN digraphs obtained by HPP or UBP which are isomorphic to a fixed weakly connected digraph. As special cases, we consider three such quantities, namely, the number of vertices with fixed indegree, the number of shared  $k$ NN pairs and the

number of reflexive  $k$ NN pairs.

In a digraph, *indegree* (resp. *outdegree*) of a vertex  $v$  is the number of arcs with head (resp. tail)  $v$  and denoted as  $d_{in}(v)$  (resp.  $d_{out}(v)$ ). Notice that the outdegree of a vertex in  $D_{nn}^{[k]}(V)$  is always  $k$  as long as  $V$  has at least  $k + 1$  vertices. For  $j \geq 0$ , let  $Q_j^{[k]}(V)$  denote the number of vertices of  $D_{nn}^{[k]}(V)$  with indegree  $j$ , that is, the number of points which are  $k$ NN of exactly  $j$  points in  $V$ . In Figure 1.1, note that  $Q_0^{[1]}(V) = 3, Q_1^{[1]}(V) = 4, Q_2^{[1]}(V) = 3$  and  $Q_j^{[1]}(V) = 0$  for every  $j \geq 3$ , and  $Q_0^{[2]}(V) = 1, Q_1^{[2]}(V) = 3, Q_2^{[2]}(V) = 2, Q_3^{[2]}(V) = 3, Q_4^{[2]}(V) = 1$  and  $Q_j^{[2]}(V) = 0$  for every  $j \geq 5$ . The problem of finding the probability that a random point is the NN of precisely  $j$  other points is studied by many authors such as Clark and Evans (1955), Roberts (1969), Newman et al. (1983) and Henze (1987). The quantities  $Q_j^{[1]}$ 's are used in tests for spatial symmetry (see, Ceyhan (2014)). Also in Enns et al. (1999),  $Q_0^{[1]}, Q_1^{[1]}$  and  $Q_2^{[1]}$  correspond to the number of “lonely”, “normal” and “popular” individuals in a population, respectively.

A pair of arcs in  $D_{nn}^{[k]}(V)$  is called *shared  $k$ NN* whenever they have a common head, and the number of shared  $k$ NNs in  $D_{nn}^{[k]}(V)$  is denoted as  $Q^{[k]}(V)$ . Note that the quantity  $Q^{[k]}$  can be expressed in terms of  $Q_j^{[k]}$ 's. By a simple double counting argument, one can easily see that

$$Q^{[k]}(V) = \sum_{v \in V} \binom{d_{in}(v)}{2} = \sum_{j \geq 0} \frac{j(j-1)}{2} Q_j^{[k]}(V)$$

for any point set  $V$ . Notice that  $Q^{[1]}(V) = 3$  and  $Q^{[2]}(V) = 17$  in Figure 1.1.

A pair of vertices  $\{u, v\}$  is called a *reflexive  $k$ NN* pair whenever  $u \in kNN(v)$  and  $v \in kNN(u)$ , that is,  $u$  and  $v$  are  $k$ NN of each other (Cox (1981)). We denote the number of reflexive pairs in  $D_{nn}^{[k]}(V)$  as  $R^{[k]}(V)$ . Notice that  $R^{[1]}(V) = 3$  and  $R^{[2]}(V) = 6$  in Figure 1.1. Some authors have called these pairs as isolated nearest neighbors (Pickard (1982)) or mutual nearest neighbors (Schilling (1986b)). In graph theory, reflexive pairs are also referred to as symmetric arcs (Chartrand and Lesniak (1996)). The quantity  $R^{[k]}$  could be of interest for inferential purposes as well, since it is a measure of mutual (symmetric) spatial dependence between points, which might indicate a special and/or stronger form of clustering of data points. For instance, a simple test based on the proportion of the number of reflexive pairs to the sample size was presented by Dacey (1960) to interpret the degree of regularity or clustering of the locations of towns alongside a river.

Numbers of reflexive and shared  $k$ NN pairs are of importance in various fields. For example, in spatial data analysis, the distributions of the tests based on nearest neighbor contingency tables depend on these two quantities (Cuzick and Edwards (1990), Dixon (1994) and Ceyhan (2009)). Moreover, neighbor sharing type quantities such as  $Q^{[1]}$  are also of interest for the problem of estimating the intrinsic dimension of a data set or cloud (see, Brito et al. (2013)).

Using the ideas and results in Penrose and Yukich (2001, 2002), we provide a general

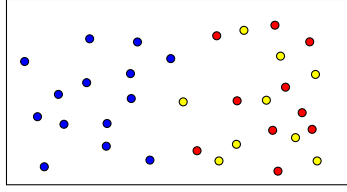


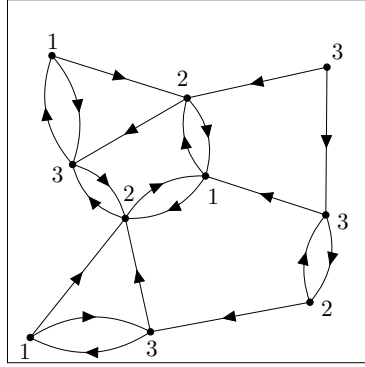
Figure 1.2: Segregation of blue points and association of yellow and red points.

result for the asymptotic distributions of the quantities based on subdigraph counts in random  $k$ NN digraphs, and as corollaries, we obtain asymptotic results for the number of vertices with a given indegree  $Q_j^{[k]}$ , the number of shared pairs  $Q^{[k]}$  and the number of reflexive pairs  $R^{[k]}$ .

A major application of  $k$ NN digraphs is in spatial data analysis. The spatial relationship between two species, groups or classes can be of interest in geography, ecology, astronomy and biology. One major question is whether two classes are spatially segregated (members of a class occur near the same class) or associated (members of a class occur near members from other classes), or neither. (See Figure 1.2 for an illustration.) Based on NN relations, many tests of spatial segregation and association have been proposed in literature, such as Pielou's NNCT Pielou (1961), Ripley's K-function Ripley (1981), Friedman and Rafsky (1983), Cuzick and Edwards (1990), Dixon (1994, 2002) and Ceyhan (2009). Among these, Cuzick and Edwards (1990) provide tests based on  $k$ NN relations for the sample consisting of two classes, whereas Dixon (2002) introduces tests based on 1NN relations for multi-class samples consisting of more than two classes. Dixon (2002) analyze the 1NN contingency table (introduced by Pielou (1961)) and applies the methods in ecology, Cuzick and Edwards (1990) consider various quantities for testing e.g. disease clustering. We generalize the idea of these two papers and provide statistics to test spatial clustering for multi-class case based on  $k$ NN relations among the sample points.

Let  $m \geq 2$  be an integer and  $\mathcal{X}_1, \dots, \mathcal{X}_m$  be pairwise disjoint classes of points in  $\mathbb{R}^d$ . We call  $\mathcal{X}_j$  as class  $j$  for each  $j \in \{1, \dots, m\}$ . For each  $i, j \in \{1, \dots, m\}$ , we are interested in the number of pairs  $\{v, u\}$  in the data such that  $v$  is from class  $i$ ,  $u$  is from class  $j$  and  $u$  is a  $k$ NN of  $v$ , denoted  $N_{ij} := N_{ij}^{[k]}$ . For each point, mark it with the label of the class it belongs to. Then,  $N_{ij}$  is the number of arcs in the  $k$ NN digraph of the union of all classes whose tail is marked with  $i$  and head is marked with  $j$ . We can classify these counts according to base and  $k$ NN classes and present a contingency table called  $k$ NNCT. In this  $m \times m$  table, the entry at the intersection of the  $i$ -th row and  $j$ -th column is  $N_{ij}$  for all  $i, j \in \{1, \dots, m\}$ . An example of a  $k$ NNCT for  $k = 2$  and  $m = 3$  is presented in Figure 1.3.

Distribution of the cell counts in a  $k$ NNCT differs with respect to the randomness



| 2NNCT of $V$ |         |
|--------------|---------|
| Classes      | Classes |
|              | 1 2 3   |
| 1            | 0 4 2   |
| 2            | 2 0 4   |
| 3            | 3 4 1   |

Figure 1.3: 2NNCT of  $V = V_1 \cup V_2 \cup V_3$  with  $V_1 = \{v_1, v_2, v_3\}$ ,  $V_2 = \{v_4, v_5, v_6\}$  and  $V_3 = \{v_7, v_8, v_9, v_{10}\}$  where  $v_1, \dots, v_{10}$  are given in Figure 1.1.

in the underlying structure. Randomness on the data can reside in the allocation of the points and/or in the assignment of the marks/labels. In our analysis, we mainly consider four random processes. In two of them, data points are obtained from HPP and UBP, and the marks are randomly assigned over the points, namely, marked HPP and marked UBP, respectively. In the other two, data points are assumed to be fixed and class labels are randomly assigned to the points and both are denoted *random labelling* (RL). With respect to whether class sizes are random or fixed we call these processes type 1 RL and type 2 RL, respectively. For each random process, we study the asymptotic behavior of the joint distribution of  $N_{ij}$ 's. Notice that diagonal cell counts in a  $k$ NNCT can be used to test spatial segregation whereas off-diagonal cell entries are convenient in testing spatial association. By the results we obtain, overall segregation (and association) tests can be used to detect any deviation from the null case of interest.

We also present a classification of random digraphs so as to determine the class or category in which random  $k$ NN digraphs fall. A random digraph is called *isomorphism-invariant* whenever isomorphic digraphs have the same probability. We classify isomorphism-invariant random digraphs according to where randomness resides, namely, arcs, vertices, and vertices and arcs together which in turn yield arc random digraphs (ARD), vertex random digraphs (VRD) and vertex-arc random digraphs (VARD), respectively. This digraph classification can be viewed as an extension of the classification of isomorphism-invariant random graphs of Beer et al. (2011). We introduce randomness in the direction of the edges of a given graph and obtain direction random digraphs (DRD) as well. We classify DRDs according to which component is random in addition to the direction and study the relations of DRDs with VARDs, VRDs and ARDs. We then determine the membership of random  $k$ NN digraphs with respect to these digraph families.

The text is organized as follows: The notations, definitions and some frequently used theorems are introduced in the next section. In Chapter 2, we first provide LLN and CLT type results on the subdigraph counts in random  $k$ NN digraphs. Then, as

corollaries, we apply those results to number of vertices with indegree  $j$ ,  $Q_j^{[k]}$ , number of shared  $k$ NNs,  $Q^{[k]}$ , and number of reflexive pairs,  $R^{[k]}$ . In Section 2.6, we study these three quantities for  $k = 1$  in one dimensional data, and Section 2.7 investigates the pairwise dependence of these random variables for  $k = 1$  under binomial point process. In Chapter 2.7, we analyze asymptotic distribution of the entries of  $k$ NNCT under four random processes. We mainly consider their joint distribution and also provide asymptotic results on quadratic forms based on the cell counts. Chapter 4 presents a classification of isomorphism-invariant random digraphs based on where randomness lies in the structure. We introduce ARDs, VRDs, VARDs and DRDs, and study pairwise relations of these random digraph families together with random  $k$ NN digraphs.

## 1.1 Preliminaries

### 1.1.1 Graph Theoretical Tools

A *graph*  $G$  is a finite non-empty set  $V(G)$  of elements called *vertices* together with a set  $E(G)$  of unordered pairs of vertices of  $G$  called *edges*. An edge  $\{u, v\}$  is denoted by  $uv$  for convenience in the text. A graph  $G_1$  is *isomorphic* to a graph  $G_2$  (or  $G_1$  and  $G_2$  are *isomorphic*) if there exists a bijection  $f : V(G_1) \rightarrow V(G_2)$  such that  $uv \in E(G_1)$  if and only if  $f(u)f(v) \in E(G_2)$ .

A *directed graph* (or simply *digraph*)  $D$  consists of a non-empty finite set  $V(D)$  of elements called *vertices* and a finite set  $A(D)$  of ordered pairs of distinct vertices called *arcs* (or *directed edges*). We call  $V(D)$  the vertex set and  $A(D)$  the arc set of  $D$ . We will often denote  $D$  as  $D = (V, A)$ . The above definition of a digraph implies that we allow a digraph to have arcs with the same end-vertices (for example, both  $(u, v)$  and  $(v, u)$  may be in  $A$ ). In this work, we only consider simple digraphs. That is, we do not allow *parallel* (also called *multiple*) arcs, i.e., pairs of arcs with the same tail and the same head, or *loops* (i.e., arcs whose heads and tails coincide). When parallel arcs and loops are admissible we speak of *directed pseudographs*; directed pseudographs without loops are *directed multigraphs* (Chartrand and Lesniak (1996)).

For an arc  $(u, v)$ , the vertex  $u$  is called the *tail* and the vertex  $v$  is called the *head*. The head and tail of an arc are called the *end-vertices*. A digraph  $D_1$  is a *subdigraph* of a digraph  $D_2$  if  $V(D_1) \subseteq V(D_2)$  and  $A(D_1) \subseteq A(D_2)$ . A digraph  $D_1$  is *isomorphic* to a digraph  $D_2$  (or  $D_1$  and  $D_2$  are *isomorphic*) if there is a bijection  $f : V(D_1) \rightarrow V(D_2)$  such that  $(u, v) \in A(D_1)$  if and only if  $(f(u), f(v)) \in A(D_2)$ .

A graph or a digraph is *finite* if its vertex set is finite. A  $u - v$  *path* in a graph  $G$  is a sequence of pairwise distinct vertices  $u = u_1, u_2, \dots, u_m = v$  such that  $u_i u_{i+1}$  is an edge in  $G$  for each  $1 \leq i \leq m - 1$ , and the *length* of the path is the number of edges in the path. A graph  $G$  is called *connected* if there exists a  $u - v$  path for every pair of vertices  $u$  and  $v$  in  $G$ . The *distance* between the vertices  $u$  and  $v$  of a connected graph

$G$  is the length of a shortest  $u - v$  path. A digraph is called *weakly connected* if its underlying graph is connected. For more information about graphs and digraphs see, e.g., Chartrand and Lesniak (1996).

A marked graph (resp. digraph) is a graph (resp. digraph) in which each vertex has a mark. Two marked graphs (resp. marked digraphs) are isomorphic if there exists an isomorphism between the graphs (resp. digraphs) such that the corresponding vertices under the isomorphism have the same mark.

### 1.1.2 Probabilistic Tools

For random variables  $X$  and  $Y$ , let  $\mathbf{E}(X)$ ,  $\mathbf{Var}(X)$  and  $\mathbf{Cov}(X, Y)$  denote the expectation of  $X$ , the variance of  $X$  and the covariance of  $X$  and  $Y$ , respectively.

Let  $(X_n)_{n \geq 1}$  be a sequence of random variables and  $X$  be a random variable. If

$$\lim_{n \rightarrow \infty} \mathbf{E}(X_n) = \mathbf{E}(X),$$

then we write  $X_n \xrightarrow{c.m.} x$  as  $n \rightarrow \infty$  (*convergence of means*). Notice that convergence of means is different than *convergence in mean* ( $(X_n)_{n \geq 1}$  converges in mean to  $X$  if  $\lim_{n \rightarrow \infty} \mathbf{E}(|X_n - X|) = 0$ ), and convergence in mean implies convergence of means but not vice versa. If

$$P(\lim_{n \rightarrow \infty} X_n = X) = 1,$$

then we say  $(X_n)_{n \geq 1}$  *almost surely* converges to  $X$  and write  $X_n \xrightarrow{a.s.} X$  as  $n \rightarrow \infty$ . If

$$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty \text{ for every } \epsilon > 0,$$

then we say  $(X_n)_{n \geq 1}$  *converges completely* to  $X$  and denote  $X_n \xrightarrow{c.c.} x$  as  $n \rightarrow \infty$ . Notice that complete convergence implies almost sure convergence but not vice versa. We use the notation  $\xrightarrow{c.m.c.c.}$ , if both convergence of means and complete convergence hold. If

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x)$$

for every  $x \in \mathbb{R}$  at which the cumulative distribution function of  $X$  is continuous, then we say  $(X_n)_{n \geq 1}$  *converges in law* (or *converges in distribution*) to  $X$  and denote  $X_n \xrightarrow{\mathcal{L}} X$  as  $n \rightarrow \infty$ .

It is also straightforward to extend convergence in distribution to random vectors. Let  $(\mathbf{X}_n = (X_{n,1}, \dots, X_{n,m}))_{n \geq 1}$  be a sequence of random vectors and  $\mathbf{X} = (X_{0,1}, \dots, X_{0,m})$  be a random vector. If

$$\lim_{n \rightarrow \infty} P(X_{n,1} \leq x_1, \dots, X_{n,m} \leq x_m) = P(X_{0,1} \leq x_1, \dots, X_{0,m} \leq x_m)$$

for any  $\mathbf{x} = (x_1, \dots, x_m) \in \mathbb{R}^m$  at which the cumulative distribution function of  $\mathbf{X}$  is continuous, then, similarly, we say  $(\mathbf{X}_n)_{n \geq 1}$  converges in law to  $\mathbf{X}$  and write  $\mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{X}$  as  $n \rightarrow \infty$ .

**Theorem 1.1.1** (Cramér-Wold Theorem).  $\mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{X}$  if and only if  $\mathbf{a}^t \mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{a}^t \mathbf{X}$  for all  $\mathbf{a} \in \mathbb{R}^m$ .

This theorem is also called as Cramér-Wold device. This shows us in what sense the notion of multivariate convergence in law is determined by the univariate notion. The multivariate convergence in law  $\mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{X}$  holds if and only if any linear combination of the components of  $\mathbf{X}_n$  converges in law to the same linear combination of the components of  $\mathbf{X}$ .

**Theorem 1.1.2** (The Continuous Mapping Theorem). For a continuous function  $g$ , if  $\mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{X}$  as  $n \rightarrow \infty$ , then  $g(\mathbf{X}_n) \xrightarrow{\mathcal{L}} g(\mathbf{X})$  as  $n \rightarrow \infty$ .

Let  $\mathcal{N}(\mu, \sigma^2)$  denote the normal distribution with mean  $\mu$  and variance  $\sigma^2$  and  $\mathcal{N}(0, 1)$  is called *standard normal variable*. Sum of squares of  $m$  independent standard normal variables is called *chi-squared variable with  $m$  degrees of freedom*, and denoted  $\chi_m^2$ .

For a random vector  $\mathbf{X} = (X_1, \dots, X_m)$ , the *covariance matrix* of  $\mathbf{X}$  is the  $m \times m$  matrix whose entry at the  $i$ -th row and the  $j$ -th column is  $\text{Cov}(X_i, X_j)$  for all  $i, j \in \{1, \dots, m\}$ , and denoted  $\text{Cov}(\mathbf{X})$ . Let  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  stand for the multivariate normal variable with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ .

**Theorem 1.1.3.** If  $\mathbf{Z} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and rank of  $\boldsymbol{\Sigma}$  is  $m$ , then

$$(\mathbf{Z} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^+ (\mathbf{Z} - \boldsymbol{\mu})$$

has a chi-squared distribution with  $m$  degrees of freedom, where  $\boldsymbol{\Sigma}^+$  stands for the MoorePenrose pseudoinverse of  $\boldsymbol{\Sigma}$ .

For the proof, see Theorem 6.2.5 in Albert (1972). Notice that if  $\boldsymbol{\Sigma}$  is invertible, then  $\boldsymbol{\Sigma}^+ = \boldsymbol{\Sigma}^{-1}$ . By combining the continuous mapping theorem and Theorem 1.1.3 we obtain the following result which is frequently used in the text.

**Corollary 1.1.4.** If a sequence of random vectors  $(\mathbf{X}_n)_{n \geq 1}$  converges in law to  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and the rank of  $\boldsymbol{\Sigma}$  is  $m$ , then we have

$$(\mathbf{X}_n - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^+ (\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{L}} \chi_m^2$$

as  $n \rightarrow \infty$ .

For any Lebesgue measurable set  $A \subset \mathbb{R}^d$ , let  $m(A)$  denote the Lebesgue measure of the set  $A$ ,  $\partial A$  denote the boundary of  $A$  and  $\text{card}(A)$  denote the cardinality of  $A$ .

A homogeneous Poisson process with intensity  $\lambda$  on entire  $\mathbb{R}^d$ , denoted  $\mathcal{P}_{(\lambda)}$ , is a point process satisfying the following two properties:

- (i) Number of points in a region  $B$ , (i.e.,  $\text{card}(\mathcal{P}_{(\lambda)} \cap B)$ ) has a Poisson distribution with mean  $\lambda \cdot m(B)$ .
- (ii) For any collection of pairwise disjoint regions  $B_1, \dots, B_s$ ,  $\text{card}(\mathcal{P}_{(\lambda)} \cap B_1), \dots, \text{card}(\mathcal{P}_{(\lambda)} \cap B_s)$  are independent.

Let  $B$  be a bounded region with positive measure, and  $U_1, \dots, U_n$  be i.i.d. (independent and identically distributed) random points which are uniformly distributed in  $B$ . Hence, the probability density of  $U_i$  is  $\mathbf{1}_{\{x \in B\}} \cdot 1/m(B)$ , where  $\mathbf{1}_{\{\cdot\}}$  is the indicator function. Note that for any region  $A \subset B$ , the number of points in  $A$  has a binomial distribution with  $p = m(A)/m(B)$ . We call such a process as *uniform binomial process* (UBP).

Let  $(\mathcal{K}, \mathcal{F}, P)$  be a probability space. In a marked point process with mark space  $(\mathcal{K}, \mathcal{F}, P)$ , independently of other points and marks, each point is assigned a mark taking value in  $\mathcal{K}$  under the distribution  $P$ . In other words, marks of the points are i.i.d. with distribution  $P$ .

## Chapter 2

SUBDIGRAPH COUNTS IN RANDOM  $k$ NN DIGRAPHS

## 2.1 Introduction

In this chapter, we study the number of vertices with fixed indegree, the number of shared  $k$ NN pairs and the number of reflexive  $k$ NNs in a  $k$ NN digraph. Using the results of Penrose and Yukich (2001, 2002), we provide LLN and CLT results on the number of copies of weakly connected digraphs in  $k$ NN digraphs under HPP and UBP. As corollaries of these results, we obtain asymptotic results for  $R^{[k]}$ ,  $Q_j^{[k]}$  and  $Q^{[k]}$ .

Let  $D$  be a fixed weakly connected digraph. For any finite point set  $V$  in  $\mathbb{R}^d$ , let  $H_D(V)$  denote the number of subdigraphs of  $D_{nn}^{[k]}(V)$  isomorphic to  $D$ . In other words,  $H_D$  counts the copies of  $D$  in the  $k$ NN digraph of the given point set. For example, if  $D$  is the digraph with  $V(D) = \{1, 2\}$  and  $A(D) = \{(1, 2), (2, 1)\}$ , then we have  $R^{[k]}(V) = H_D(V)$ . Similarly, we have  $Q^{[k]}(V) = H_D(V)$  whenever  $D$  is the digraph with  $V(D) = \{1, 2, 3\}$  and  $A(D) = \{(1, 2), (3, 2)\}$ .

$Q_j^{[k]}$  itself is not a subdigraph count but it is a linear combination of such quantities. By space packing properties of  $\mathbb{R}^d$ , there exists a constant  $K$  (depending only on  $d$  and  $k$ ) such that any member of a finite point set in  $\mathbb{R}^d$  is a  $k$ NN of at most  $K$  other members. The constant  $K$  is studied in detail in Section 2.3. Now let  $D_i$  be the digraph with  $V(D_i) = \{1, 2, \dots, i+1\}$  and  $A(D_i) = \{(1, i+1), (2, i+1), \dots, (i, i+1)\}$ , for every  $1 \leq i \leq K$ . Also, let  $D_0$  be the digraph with only one vertex, i.e.,  $V(D_0) = \{1\}$  and  $A(D_0) = \emptyset$ . Notice that  $H_{D_0}(V) = \text{card}(V)$  for any finite point set  $V$ . Then, by principle of inclusion exclusion, it is easy to see that

$$Q_j^{[k]}(V) = \sum_{i=j}^K (-1)^{i-j} \binom{i}{j} H_{D_i}(V), \quad (2.1)$$

for every  $0 \leq j \leq K$  and finite point set  $V$ .

Fix a bounded region  $B_0 \subset \mathbb{R}^d$  such that  $m(B_0) > 0$  and  $m(\partial(B_0)) = 0$ . Let  $\mathcal{P}_n$  denote the HPP of intensity  $n/m(B_0)$  restricted to  $B_0$  and  $\mathcal{U}_n = \{U_1, \dots, U_n\}$  denote the UBP over  $B_0$ . Throughout this thesis, the random point sets we consider are  $\mathcal{P}_n$  and  $\mathcal{U}_n$ .

**Theorem 2.1.1** (Main Theorem). *Let  $m$  be a given positive integer,  $D_1, \dots, D_m$  be finite weakly connected digraphs,  $a_1, \dots, a_m$  be real numbers and  $H(V) = a_1 H_{D_1}(V) + \dots + a_m H_{D_m}(V)$  for every finite  $V \subset \mathbb{R}^d$ . Then there exist constants  $\xi, \tau^2, \sigma^2$  with*

$0 \leq \tau^2 \leq \sigma^2$  such that as  $n \rightarrow \infty$ ,

$$\begin{aligned} n^{-1}H(\mathcal{U}_n) &\xrightarrow{c.m.c.c.} \xi, \\ n^{-1}\mathbf{Var}(H(\mathcal{U}_n)) &\rightarrow \tau^2, \\ n^{-1/2}(H(\mathcal{U}_n) - \mathbf{E}(H(\mathcal{U}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau^2), \\ n^{-1}H(\mathcal{P}_n) &\xrightarrow{c.m.c.c.} \xi, \\ n^{-1}\mathbf{Var}(H(\mathcal{P}_n)) &\rightarrow \sigma^2, \\ n^{-1/2}(H(\mathcal{P}_n) - \mathbf{E}(H(\mathcal{P}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2). \end{aligned}$$

Moreover,  $\xi$ ,  $\tau^2$  and  $\sigma^2$  are independent of the choice of the support  $B_0$ .

A crucial condition on the given digraphs is connectedness; and another is number of vertices of the subdigraph should be fixed. The proof of Theorem 2.1.1 is given in the next section.

## 2.2 Proof of the Main Theorem

We first borrow some notation and definitions regarding the CLT and LLN results from Penrose and Yukich (2001, 2002).

For any set  $A \subset \mathbb{R}^d$  and  $y \in \mathbb{R}^d$ , denote by  $A + y$  the set  $\{x + y : x \in A\}$ . Also, for any  $c \in \mathbb{R}$ , let  $cA$  denote the set  $\{cx : x \in A\}$ . For  $x \in \mathbb{R}^d$  and  $r > 0$ , let  $B_r(x)$  denote the Euclidean open ball centered at  $x$  and with radius  $r$ . Let  $\text{card}(A)$  denote the cardinality of  $A$ . Let  $\mathcal{P}$  be HPP of unit intensity on  $\mathbb{R}^d$ .

Let  $H$  be a real valued functional defined for all finite subsets of  $\mathbb{R}^d$ .  $H$  is called *translation-invariant*, if  $H(V + y) = H(V)$  for all  $V \subset \mathbb{R}^d$  and  $y \in \mathbb{R}^d$ , and *scale-invariant*, if  $H(cV) = H(V)$  for all  $V \subset \mathbb{R}^d$  and  $c \neq 0$ . Notice that the functionals we consider only depend on the ordering of the pairwise distances between sample points. Henceforth, these functionals are translation-invariant and scale-invariant, and thus, without loss of generality, we may assume that  $B_0$  is of unit volume and contains the origin.

We say  $H$  is *linearly bounded*, if there is a constant  $c_1$  such that  $|H(V)| \leq c_1 \cdot \text{card}(V)$  for every finite  $V \subset \mathbb{R}^d$ . Remaining conditions on  $H$  are defined in terms of the *add one cost*, meaning that the increment in  $H$  caused by inserting a point at the origin into a finite point set  $V \subset \mathbb{R}^d$ , which is formally given by

$$\Delta_H(V) := H(V \cup \{0\}) - H(V).$$

The functional  $H$  has *bounded add one cost*, if there exists a constant  $c_2$  such that  $|\Delta_H(V)| \leq c_2$  for every finite point set  $V \subset \mathbb{R}^d$ .

An almost surely finite random variable  $S$  is called a *radius of stabilization* of  $H$ ,

if the add one cost is unaffected by the changes in the configuration outside the ball  $B_S(0)$ ; that is,

$$\Delta_H((\mathcal{P} \cap B_S(0)) \cup A) = \Delta_H(\mathcal{P} \cap B_S(0))$$

for every finite  $A \subset \mathbb{R}^d \setminus B_S(0)$ , and in this case, this add one cost (i.e.,  $\Delta_H(\mathcal{P} \cap B_S(0))$ ) is denoted by  $\Delta_H(\infty)$ . Note that  $\Delta_H(\infty)$  is a random variable as well.

### 2.2.1 CLT Results

One can easily obtain the following proposition by applying Theorem 2.1 in Penrose and Yukich (2001) (see, Appendix A.1).

**Proposition 2.2.1.** *Suppose that  $H$  is translation-invariant, scale-invariant, linearly bounded, has a radius of stabilization and bounded add one cost. Then there exist constants  $\sigma^2$  and  $\tau^2$ , with  $0 \leq \tau^2 \leq \sigma^2$ , such that as  $n \rightarrow \infty$ ,*

$$\begin{aligned} \mathbf{Var}(H(\mathcal{P}_n))/n &\rightarrow \sigma^2, \\ \mathbf{Var}(H(\mathcal{U}_n))/n &\rightarrow \tau^2, \\ n^{-1/2}(H(\mathcal{P}_n) - \mathbf{E}(H(\mathcal{P}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2), \\ n^{-1/2}(H(\mathcal{U}_n) - \mathbf{E}(H(\mathcal{U}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau^2). \end{aligned}$$

Also,  $\sigma^2$  and  $\tau^2$  are independent of the choice of the support  $B_0$ . Moreover, if the distribution of  $\Delta_H(\infty)$  is non-degenerate, then  $\tau^2 > 0$  and hence also  $\sigma^2 > 0$ .

**Lemma 2.2.2.** *If  $H_i$  satisfies the conditions of Proposition 2.2.1 for each  $1 \leq i \leq m$ , so does  $a_1 H_1 + \dots + a_m H_m$  for any real numbers  $a_1, \dots, a_m$ .*

*Proof.* First note that by applying induction on  $m$ , it suffices to prove the claim for  $m = 2$ . Let  $H = a_1 H_1 + a_2 H_2$ .

$$H(V + y) = a_1 H_1(V + y) + a_2 H_2(V + y) = a_1 H_1(V) + a_2 H_2(V) = H(V)$$

for all  $V \subset \mathbb{R}^d$  and  $y \in \mathbb{R}^d$  since  $H_1$  and  $H_2$  are translation-invariant, and hence  $H$  is translation-invariant as well.

As  $H_1$  and  $H_2$  are scale-invariant, for all  $V \subset \mathbb{R}^d$  and every nonzero  $a \in \mathbb{R}$ , we have

$$H(aV) = a_1 H_1(aV) + a_2 H_2(aV) = a_1 H_1(V) + a_2 H_2(V) = H(V)$$

which implies that  $H$  is scale-invariant too.

As both  $H_1$  and  $H_2$  are linearly bounded and have bounded add one costs, there exist constants  $c_1$  and  $c_2$  such that  $|H_i(V)| \leq c_i \cdot \text{card}(V)$  and  $|\Delta_{H_i}(V)| \leq c_i$ , for  $i = 1, 2$ .

So, by triangular inequality, we have

$$|H(V)| = |a_1 H_1(V) + a_2 H_2(V)| \leq (|a_1|c_1 + |a_2|c_2) \cdot \text{card}(V),$$

which implies that  $H$  is linearly bounded.

Let  $S_i$  be a radius of stabilization of  $H_i$  for  $i = 1, 2$ , and  $S = \max\{S_1, S_2\}$ . Then,  $S$  is a radius of stabilization for both  $H_1$  and  $H_2$ . Since

$$\begin{aligned} \Delta_H(V) &= H(V \cup \{0\}) - H(V) = a_1 H_1(V \cup \{0\}) + a_2 H_2(V \cup \{0\}) - (a_1 H_1(V) + a_2 H_2(V)) \\ &= a_1 \Delta_{H_1}(V) + a_2 \Delta_{H_2}(V) \end{aligned} \quad (2.2)$$

for any finite point set  $V \subset \mathbb{R}^d$ , we have  $\Delta_H(\infty) = a_1 \Delta_{H_1}(\infty) + a_2 \Delta_{H_2}(\infty)$  and thus,  $S$  is a radius of stabilization for  $H$ .

By triangular inequality and Equation (2.2), we obtain

$$|\Delta_H(V)| \leq |a_1| |\Delta_{H_1}(V)| + |a_2| |\Delta_{H_2}(V)| \leq |a_1|c_1 + |a_2|c_2,$$

for any  $V \subset \mathbb{R}^d$ , so  $H$  has bounded add one cost, and so the result follows.  $\square$

Let  $D$  be a finite weakly connected digraph. Then by Lemma 2.2.2 it suffices to show that  $H_D$  satisfies the conditions of Proposition 2.2.1 to prove the CLT results in Theorem 2.1.1.

Note that  $D_{nn}^{[k]}(V)$  depends only on the ordering of the pairwise distances of the points in  $V$ , and thus, one can easily see that  $H_D$  is translation-invariant and scale-invariant.

For any point  $v$  in  $V$  let  $h_D(V, v)$  denote the number of copies of  $D$  in  $D_{nn}^{[k]}(V)$  containing  $v$ . By Lemma 2.3.2 we have  $d_{in}(v) \leq k\kappa'(d)$  where  $\kappa'(d)$  is a constant which only depends on the dimension  $d$  ( $\kappa'(d)$  is defined in the Section 2.3.1). We also have  $d_{out}(v) = k$ , and therefore,  $v$  is adjacent to at most  $K = k(\kappa'(d) + 1)$  arcs in  $D_{nn}^{[k]}(V)$ . Let  $s$  be the number of vertices in  $D$  i.e.,  $s = \text{card}(V(D))$ . Then as  $D$  is weakly connected, it is easy to verify that

$$0 \leq h_D(V, v) \leq C, \quad (2.3)$$

for some constant  $C := C(K, s)$  which only depends on  $K$  and  $s$  (i.e.,  $C$  is independent of  $V$ ). Since each copy of  $D$  in  $D_{nn}^{[k]}(V)$  has exactly  $s$  vertices we get

$$sH_D(V) = \sum_{v \in V} h_D(V, v). \quad (2.4)$$

Thus,  $|H_D(V)| \leq (C/s) \cdot \text{card}(V)$  and so,  $H_D$  is linearly bounded.

Now suppose that the point 0 at the origin is not in  $V$ . Inserting the point 0 to  $V$  may cause addition or deletion of some arcs in the  $k$ NN digraph. We definitely add arcs

$(0, v)$  where  $v$  is a  $k$ NN of 0 in  $V \cup \{0\}$ . Also some arcs of the form  $(v, 0)$  are inserted whenever 0 is a  $k$ NN to  $v$ . Notice that if  $u$  is not a  $k$ NN of  $v$ , it is still not a  $k$ NN to  $v$  after the insertion of 0. Clearly, an arc  $(v, u)$  is deleted after the addition of 0 only if  $u$  is the  $k$ -th NN of  $v$  in  $V$  and  $\|v - 0\| < \|v - u\|$ . And in this case, 0 becomes a  $k$ NN of  $v$ . Let  $v_1, \dots, v_p$  be points adjacent to 0 in  $D_{nn}^{[k]}(V \cup \{0\})$  (i.e.,  $(0, v_i)$  or  $(v_i, 0)$  is an arc in the  $k$ NN digraph after the insertion of 0) and recall that  $p \leq K$ . Then any deleted or created copy of  $D$  in the  $k$ NN digraph contains at least one of the  $v_i$ 's. Since there are at most  $C$  copies of  $D$  containing a given vertex, we see that  $|\Delta_{H_D}(V)| \leq pC \leq KC$  and hence,  $H_D$  has a bounded add one cost.

We finally show that  $H_D$  has a radius of stabilization. But, we first construct a setting essential for the proof. We show that there exist cones  $C_1, \dots, C_m$  with 0 as their common peak such that  $x, y \in C_i \setminus \{0\}$  implies  $\|x - y\| < \max\{\|x\|, \|y\|\}$  for all  $1 \leq i \leq m$ , and  $\cup_{i=1}^m C_i = \mathbb{R}^d$  (see, Lemma S in Appendix of Bickel and Breiman (1983)).

Note that the union of the open balls  $B_{1/2}(x)$  where  $x \in \partial B_1(0)$  is an open covering of  $\partial B_1(0)$ . Since  $\partial B_1(0)$  is compact, there exists a finite subcover, say  $B_{1/2}(x_1), \dots, B_{1/2}(x_m)$ . Let  $B'_i = B_{1/2}(x_i) \cap \partial B_1(0)$  and  $C_i = \{ax : x \in B'_i, a \geq 0\}$  for all  $1 \leq i \leq m$ . Recall that whenever  $x, y \in B'_i$ , by the triangular inequality, we have  $\|x - y\| < \|x - x_i\| + \|x_i - y\| \leq 1/2 + 1/2 = 1$ . Now if  $x, y \in C_i \setminus \{0\}$ , then  $x = ax'$  and  $y = by'$  for some  $a, b > 0$  and  $x', y' \in B'_i$ . Assume that  $a \leq b$ . Then, we have

$$\|x - y\| = \|a(x' - y') - (b - a)y'\| \leq a\|x' - y'\| + (b - a)\|y'\| < a + (b - a) = b = \|y\|$$

which shows  $\|x - y\| < \max\{\|x\|, \|y\|\}$ . One can also obtain this result by a geometric argument as follows: By construction, the angle  $\widehat{x0y}$  is less than  $60^\circ$  and therefore, the edge  $[xy]$  is not the largest edge of the triangle  $\triangle(x0y)$ , that is,  $\|x - y\| < \max\{\|x\|, \|y\|\}$ . Moreover, it is easy to see that the union of the cones is the whole space since  $x/\|x\| \in B'_i$  for some  $i$  and hence  $x \in C_i$ , for each nonzero  $x \in \mathbb{R}^d$ .

Let  $C_i(t) = \{ax : x \in B'_i, 0 \leq a \leq t\}$  for all positive integers  $t$  and  $1 \leq i \leq m$ . Given the HPP  $\mathcal{P}$  of intensity 1 on  $\mathbb{R}^d$ , let the random variable  $T$  be the minimum  $t$  such that each cone  $C_i(t)$  contains at least  $k + 1$  points from  $\mathcal{P}$ , and set  $S_0 = T + 1$ . Then  $S_0$  is a.s. finite, since  $C_i = \cup_{t=1}^\infty C_i(t)$  contains infinitely many points from  $\mathcal{P}$  almost surely for each  $i$ .

**Lemma 2.2.3.** *Let  $v$  be a nonzero point in  $V = (\mathcal{P} \cap B_{S_0}(0)) \cup A \cup \{0\}$  for some finite  $A \subset \mathbb{R}^d \setminus B_{S_0}(0)$ . In  $V$ , if  $0 \in kNN(v)$  or  $v \in kNN(0)$ , then  $\|v\| < S_0$ .*

*Proof.* First note that  $v$  is in  $C_i$  for some  $i$ , and by definition of  $S_0$ , there exist points  $u_1, \dots, u_{k+1}$  in  $V$  lying in  $C_i(S_0)$ . We prove the claim by contrapositive. Suppose that  $\|v\| > S_0$ . Then,  $\|u_j - 0\| < S_0 < \|v - 0\|$  for every  $1 \leq j \leq k + 1$ , and hence  $v$  is not a  $k$ NN of 0. Moreover, by the construction of  $C_i$  and since  $\|u_j\| < S_0 < \|v\|$  for

each  $1 \leq j \leq k+1$ , we have  $\|v - u_j\| < \max\{\|u_j\|, \|v\|\} = \|v\| = \|v - 0\|$  for every  $1 \leq j \leq k+1$ . Therefore, 0 is not a  $k$ NN to  $v$ .  $\square$

We inductively construct  $S_0, S_1, S_2, \dots$  as follows: Let the random variable  $T_j$  be the minimum  $t$  such that each slice  $C_i(t) \setminus C_i(3S_j)$  contains at least  $k$  points from  $\mathcal{P}$  and set  $S_{j+1} = 3T_j$ . Note that by similar arguments for  $S_0$ , each  $S_j$  is clearly a.s. finite. Also notice that  $S_{j+1} > 9S_j$  and  $T_j > 3S_j$  for every  $j$ .

**Lemma 2.2.4.** *Let  $u$  and  $v$  be points in  $V = (\mathcal{P} \cap B_{S_{j+1}}(0)) \cup A$  for some finite  $A \subset \mathbb{R}^d \setminus B_{S_{j+1}}(0)$  and suppose  $\|u\| < S_j$ . If  $u \in kNN(v)$  or  $v \in kNN(u)$ , then  $\|v\| < S_{j+1}$ .*

*Proof.* The proof is again by contrapositive. Assume  $\|v\| > S_{j+1}$ .

As  $v$  is in  $C_i$  for some  $i$  and by construction of  $S_{j+1}$ , there exist points  $v_1, \dots, v_k$  in  $C_i(T_j) \setminus C_i(3S_j)$  from  $V$ . Triangular inequality implies

$$\|u - v_s\| \leq \|u\| + \|v_s\| < S_j + T_j \quad (2.5)$$

for every  $1 \leq s \leq k$ , and

$$\|u - v\| \geq \|v\| - \|u\| > \|v\| - S_j > 3T_j - S_j. \quad (2.6)$$

Since  $T_j > 3S_j$ , we have  $3T_j - S_j > S_j + T_j$ , and thus, inequalities in (2.5) and (2.6) yield  $\|u - v_s\| < \|u - v\|$  for each  $1 \leq s \leq k$ . Therefore,  $v$  is not a  $k$ NN of  $u$ .

By (2.6) we have

$$\|u - v\|^2 > (\|v\| - S_j)^2 = \|v\|^2 - 2S_j\|v\| + S_j^2 > \|v\|^2 - 2S_j\|v\|. \quad (2.7)$$

Notice that  $\|v_s\| > 3S_j$  and  $\|v\| > S_{j+1} = 3T_j > 3\|v_s\|$  which implies  $\|v\| - \|v_s\| > 2\|v\|/3$ . Then we get

$$\|v_s\|(\|v\| - \|v_s\|) > 3S_j \frac{2\|v\|}{3} = 2S_j\|v\|. \quad (2.8)$$

Inequalities in (2.7) and (2.8) yield

$$\|v - u\|^2 > \|v\|^2 - \|v_s\|(\|v\| - \|v_s\|) = \|v\|^2 - \|v_s\|\|v\| + \|v_s\|^2, \quad (2.9)$$

for each  $1 \leq s \leq k$ . Moreover, the construction of  $C_i$  implies  $\widehat{v0v_s} < 60^\circ$ , and hence by the cosine theorem in triangles, we have

$$\|v - v_s\|^2 < \|v\|^2 + \|v_s\|^2 - \|v\|\|v_s\|, \quad (2.10)$$

for each  $1 \leq s \leq k$ . Then, by the inequalities in (2.9) and (2.10), we obtain  $\|v - v_s\| < \|v - u\|$  for all  $1 \leq s \leq k$ . Thus,  $u$  is not a  $k$ NN of  $v$ .  $\square$

Now let  $l$  be the length of a longest path in the underlying graph of  $D$ . We claim that  $S_l$  is a radius of stabilization for  $H_D$ . Let  $V = (\mathcal{P} \cap B_{S_l}(0)) \cup A$  for some finite  $A \subset \mathbb{R}^d \setminus B_{S_l}(0)$ . Recall that, after the addition of 0, every new or disappeared copy of  $D$  contains at least one of the vertices adjacent to 0 in  $D_{nn}^{[k]}(V \cup \{0\})$ . Let  $F$  be the set of these vertices. Suppose a copy of  $D$  contains  $u \in F$  and  $v$  be another vertex of this copy. Since  $D$  is weakly connected, there exists a  $u - v$  path in the underlying graph of  $D_{nn}^{[k]}(V)$ , say  $u = u_0, u_1, \dots, u_s = v$ . Note that we have  $s \leq l$  and  $u_i \in kNN(u_{i-1})$  or  $u_{i-1} \in kNN(u_i)$  for every  $1 \leq i \leq s$ . Lemma 2.2.3 implies  $\|u_1\| < S_0$ . Then, inductively Lemma 2.2.4 gives  $\|u_i\| < S_i$  for each  $1 \leq i \leq s$  and hence,  $\|u_s\| = \|v\| < S_s \leq S_l$ , that is,  $\|v\| < S_l$ . Therefore, any change in the  $k$ NN digraph caused by the insertion of the origin occurs in the ball  $B_{S_l}(0)$  and so, the set  $A$  which is outside of this ball does not effect the add one cost. In other words,  $S_l$  is a radius of stabilization for  $H_D$ .

### 2.2.2 Convergence of Means and Complete Convergence Results

We next prove c.m.c.c. results in Theorem 2.1.1 applying Theorem 3.2 in Penrose and Yukich (2002). Their result is on the functionals of the form  $\sum_{v \in V} h(V, v)$  where  $h(V, v)$  is a functional defined for every finite point set  $V \subset \mathbb{R}^d$  and  $v \in V$ . Although their result is for marked point sets, the theorem is obviously still true for unmarked sets as the unmarked point set can be viewed as a marked point set with a single mark.

The functional  $h$  is said to be *translation-invariant* if  $h(V, v) = h(V + y, v + y)$  for every finite  $V \subset \mathbb{R}^d$ ,  $v \in V$  and  $y \in \mathbb{R}^d$ ;  $h$  is called *scale-invariant* if  $h(V, v) = h(aV, av)$  for every  $V \subset \mathbb{R}^d$ ,  $v \in V$  and  $a \in \mathbb{R} \setminus \{0\}$ . We say that  $h$  is *uniformly bounded*, if there exists a constant  $c$  such that  $|h(V, v)| \leq c$  for all  $V \subset \mathbb{R}^d$  and  $v \in V$ . The functional  $h$  is called *strongly stabilizing* if there exist a.s. finite random variables  $S$  (a *radius of stabilization* for  $h$ ) and  $h_\infty$  (the *limit* of  $h$ ) such that with probability 1,

$$h(\mathcal{P} \cap B_S(0) \cup \{0\} \cup A, 0) = h_\infty,$$

for every finite set  $A \subset \mathbb{R}^d \setminus B_S(0)$ .

By Theorem 3.2 in Penrose and Yukich (2002), one can obtain the following proposition (see, Appendix A.1.1).

**Proposition 2.2.5.** *Let  $h$  be a functional defined on pairs  $(V, v)$  consisting of a point set  $V$  in  $\mathbb{R}^d$  and an element  $v$  of  $V$ , and  $H(V) = \sum_{v \in V} h(V, v)$  for every finite  $V \subset \mathbb{R}^d$ . If  $H$  has a bounded add one cost,  $h$  is translation invariant, scale invariant, uniformly bounded and strongly stabilizing with limit  $h_\infty$ , then as  $n \rightarrow \infty$  we have*

$$n^{-1}H(\mathcal{U}_n) \xrightarrow{\text{c.m.c.c.}} \mathbf{E}(h_\infty) \text{ and } n^{-1}H(\mathcal{P}_n) \xrightarrow{\text{c.m.c.c.}} \mathbf{E}(h_\infty).$$

We now prove the c.m.c.c. results in the main theorem using Proposition 2.2.5. Since  $X_n \xrightarrow{\text{c.m.c.c.}} x$  and  $Y_n \xrightarrow{\text{c.m.c.c.}} y$  as  $n \rightarrow \infty$  implies  $aX_n + bY_n \xrightarrow{\text{c.m.c.c.}} ax + by$  as  $n \rightarrow \infty$

for any real numbers  $a$  and  $b$ , it suffices to prove c.m.c.c. results in the main theorem only for  $H_D$ . Let  $D$  be a fixed weakly connected digraph with  $s$  vertices. Let  $l$  be the length of a maximal path in the underlying graph of  $D$ . Recall that by (2.4), we have  $H_D(V) = \sum_{v \in V} h_D(V, v)/s$  for any point set  $V$ . Now set  $h = h_D/s$ . We show that  $h$  and  $H = H_D$  satisfy the conditions of Proposition 2.2.5. We have already shown that  $H_D$  has a bounded add one cost in the previous subsection. Clearly,  $h_D$  is both translation-invariant and scale-invariant, and so does  $h$ . Recall that by (2.3) we have  $0 \leq h_D(V, v) \leq C(K, s)$  for every point set  $V$  and element  $v$  of  $V$ , and hence,  $h = h_D/s$  is uniformly bounded. Finally, by the same arguments in the proof of strong stabilization of  $H_D$ , it is easy to check that (using Lemma 2.2.3 and Lemma 2.2.4)  $S_l$  is a radius of stabilization for  $h$ . Also, as  $h_D$  is uniformly bounded,  $h_\infty$  (the limit of  $h$ ) is a.s. finite and therefore,  $h$  is strongly stabilizing. So, the result follows.

**Remark 2.2.6.  $k$ NN Graphs:** A widely studied object in statistics and probability is  $k$ NN graphs (see, e.g., Friedman and Rafsky (1983), Avram and Bertsimas (1993), Penrose and Yukich (2001) and Wade (2007)).  $k$ NN graph of a point set is obtained by putting an edge between two points whenever one of them is a  $k$ NN of the other one. In other words,  $k$ NN graph is the underlying graph of the  $k$ NN digraph. The results we obtain for  $k$ NN digraphs are also valid for  $k$ NN graphs. Let  $G_1, \dots, G_m$  be finite connected graphs. For each  $1 \leq i \leq m$ , let  $H_{G_i}(V)$  denote the number of subgraphs of the  $k$ NN graph of  $V$  which are isomorphic to  $G_i$ . Then, any linear combination of  $H_{G_i}$ 's satisfies all the asymptotic properties of  $H$  in Theorem 2.1.1.

**Remark 2.2.7. Marked Point Processes:** By Theorem 3.1 of Penrose and Yukich (2002), all the asymptotic results obtained in this chapter are valid for marked point processes as well. Let  $(\mathcal{K}, \mathcal{F}, P)$  be a probability space. Recall that in a marked point process with mark space  $(\mathcal{K}, \mathcal{F}, P)$ , independently of other points and marks, every point is assigned with a mark taking value in  $\mathcal{K}$  under the distribution  $P$ . A marked graph (resp. digraph) is a graph (resp. digraph) in which each vertex has a mark. Two marked graphs (resp. marked digraphs) are isomorphic if there exists an isomorphism between the graphs (resp. digraphs) such that the corresponding vertices under the isomorphism have the same mark. Suppose  $\mathcal{U}_n$  and  $\mathcal{P}_n$  are marked UBP and HPP, respectively. Then,  $D_i$ 's in Theorem 2.1.1 and  $G_i$ 's in Remark 2.2.6 can be taken to be marked digraphs and marked graphs, respectively, and all the asymptotic results still hold. For example, fix a positive integer  $m$ . Let  $\mathcal{K} = \{1, \dots, m\}$ ,  $\mathcal{F} = 2^{\mathcal{K}}$  and  $P(\text{mark is } i) = p_i > 0$  for every  $1 \leq i \leq m$ . For every  $1 \leq i, j \leq m$ , let  $D_{ij}$  be the marked digraph with  $V(D_{ij}) = \{u, v\}$ ,  $A(D_{ij}) = \{(u, v)\}$  and marks of  $u$  and  $v$  are  $i$  and  $j$ , respectively. Let  $N_{ij}(V)$  denote the number of marked subdigraphs of  $D_{nn}^{[k]}(V)$  isomorphic to  $D_{ij}$  for any marked point set  $V$ . In other words,  $N_{ij}$  counts the arcs whose tail has mark  $i$  and head has mark  $j$  in the  $k$ NN digraph of the given marked point set. Let  $\mathcal{U}_n$  and  $\mathcal{P}_n$  be as defined before with mark space  $(\mathcal{K}, \mathcal{F}, P)$ . Then,  $H$  in Theorem 2.1.1 can be replaced with not only

$N_{ij}$  but also any linear combination of  $N_{ij}$ 's. Asymptotic behavior of  $N_{ij}$ 's under  $\mathcal{P}_n$  and  $\mathcal{U}_n$  is studied in detail in Sections 3.2 and 3.3, respectively.

As the proofs of Theorem 2.1.1 and the statements in Remarks 2.2.6 and 2.2.7 are very similar, we only gave the proof of the main theorem.

In Sections 4-6, we consider special cases of  $H_D(V)$  and apply our main theorem on them.

### 2.3 Number of Vertices with Indegree $j$

In this section, we study  $Q_j^{[k]}$ 's. We first identify the degenerate ones (i.e., those satisfy  $Q_j^{[k]}(V) = 0$  for any point set  $V$ ).

#### 2.3.1 Upper Bound on the Indegree

Let  $V$  be a finite subset of  $\mathbb{R}^d$ . We show that indegrees in  $D_{nn}^{[k]}(V)$  are bounded above by a constant which only depends on  $d$  and  $k$ .

**Definition 2.3.1.** The *kissing number* in  $\mathbb{R}^d$  is the maximum number of equal nonoverlapping spheres in  $\mathbb{R}^d$  that can touch a fixed sphere of the same size and denoted as  $\kappa(d)$ .

Currently, the exact value of  $\kappa(d)$  is known only for  $d = 1, 2, 3, 4, 8, 24$  and  $\kappa(1) = 2, \kappa(2) = 6, \kappa(3) = 12, \kappa(4) = 24, 40 \leq \kappa(5) \leq 44, 72 \leq \kappa(6) \leq 78, 126 \leq \kappa(7) \leq 134, \kappa(8) = 240, \kappa(24) = 196560$  (Musin (2008), Conway and Sloane (1988)). For general  $d$ , an asymptotic upper bound for  $\kappa(d)$  is  $2^{0.401d(1+o(1))}$  (Kabatiansky and Levenshtein (1978)) and an asymptotic lower bound is  $2^{0.2075d(1+o(1))}$  (Wyner (1965)). For more information about the kissing number problem see Conway and Sloane (1988).

For any distinct points  $x, y, z \in \mathbb{R}^d$ , let  $\triangle(xyz)$  denote the triangle with vertices  $x, y$  and  $z$ . Let  $\widehat{xyz}$  denote the angle belonging to the vertex  $y$  in  $\triangle(xyz)$  and  $[xy]$  denote the line segment (i.e., edge of  $\triangle(xyz)$ ).

The kissing number problem can be stated in another way: the maximum value of  $m$  such that there exist  $m$  points  $a_1, a_2, \dots, a_m$  lying on the unit sphere in  $\mathbb{R}^d$  such that  $\|a_i - a_j\| \geq 1$  for all  $1 \leq i \neq j \leq m$  or equivalently,  $\widehat{a_i 0 a_j} \geq 60^\circ$  for all  $1 \leq i \neq j \leq m$  where  $0$  is the origin (also the center of the unit sphere). To see that this is exactly the same problem, consider any arrangement of non-overlapping spheres touching a central sphere. By a suitable translation and scaling, we may assume that the central sphere is  $B_{1/2}(0)$  and the other balls are of radius  $1/2$ . Clearly the centers of the touching balls are on the unit sphere and the distance between any two of them is  $\geq 1$ , since they do not overlap. Additionally, for any two points  $a$  and  $b$  lying on the unit sphere, we have  $\|a - b\| \geq 1$  if and only if  $\widehat{a 0 b} \geq 60^\circ$ , since  $\triangle(a 0 b)$  is an isosceles triangle with two edges of length 1.

Let  $\kappa' := \kappa'(d)$  be the maximum value of  $m$  such that there exist  $m$  points  $a_1, a_2, \dots, a_m$  lying on the unit sphere in  $\mathbb{R}^d$  such that  $\|a_i - a_j\| > 1$  for all  $1 \leq i \neq j \leq m$ . Clearly  $\kappa'(d) \leq \kappa(d)$  and  $\kappa'(1) = 2$ . Note that if in all configurations with  $\kappa(d)$  points satisfying the kissing number problem there exist two points of distance 1, then we have strict inequality  $\kappa'(d) < \kappa(d)$ . For example,  $\kappa(2) = 6$  and it is easy to see that the only configuration with six points is the set of vertices of a regular hexagon on the unit circle. On the other hand, vertices of a regular pentagon on the unit circle is an example with five points for  $\kappa'(2)$  and hence we have  $\kappa'(2) = 5$ . We also have  $\kappa'(3) = 12$  since the vertices of a regular icosahedron on the unit ball is an example with twelve points for the kissing number problem with no pair of vertices of distance 1. For small values of  $d$ , we assert that  $\kappa'(d) \in \{\kappa(d) - 1, \kappa(d)\}$ . Currently, this assertion remains as a conjecture. The reason why we consider  $\kappa'(d)$  is explained in the following lemma.

**Lemma 2.3.2.** *Let  $v$  be a vertex of  $D_{nn}^{[k]}(V)$  where  $V$  is a random point set in  $\mathbb{R}^d$  obtained by  $\mathcal{P}_n$  or  $\mathcal{U}_n$ . Then, we have  $d_{in}(v) \leq \kappa'k$  a.s.*

*Proof.* By a convenient translation, we may suppose that  $v = 0$ . Let  $m = d_{in}(0)$  in  $D_{nn}^{[k]}(V)$ . We first prove the claim for  $k = 1$ . If  $m = 0$ , the claim follows trivially. Otherwise, there exist  $v_1, \dots, v_m \in V$  such that  $1NN(v_i) = 0$  for all  $1 \leq i \leq m$ . Therefore, for all  $1 \leq i \neq j \leq m$  we have  $\|v_i - v_j\|$  is greater than both of  $\|v_i\|$  and  $\|v_j\|$ . In other words,  $[v_i v_j]$  is the largest edge of the triangle  $\triangle(v_i 0 v_j)$ . Thus, the angle  $\widehat{v_i 0 v_j}$  is greater than  $60^\circ$ .

Now, for each  $i$  let  $a_i$  be the point on the ray  $\overrightarrow{0v_i}$  such that  $\|v_i\| = 1$ . Note that  $\widehat{a_i 0 a_j} = \widehat{v_i 0 v_j} > 60^\circ$  and hence  $[a_i a_j]$  is the largest edge of the isosceles triangle  $\triangle(a_i 0 a_j)$ . Therefore, we obtain that  $\|a_i - a_j\| > 1$  for all  $i \neq j$ , and hence  $m \leq \kappa'$  by the definition of  $\kappa'$ .

For general  $k$ , we follow the idea for  $d = 2$  introduced in Cuzick and Edwards (1990). Let  $v_1, \dots, v_m \in V$  such that  $0 \in kNN(v_i)$  for each  $1 \leq i \leq m$ . Let  $u_1$  be the  $v_i$  with largest  $\|v_i\|$  and  $r_1$  be  $\|u_1\|$ . Delete all  $v_i$ 's lying in the ball  $B_{r_1}(u_1)$ . Of the remaining  $v_i$ 's let  $u_2$  be the one with largest  $\|v_i\|$ , and  $r_2$  be  $\|u_2\|$ . Delete all  $v_i$ 's lying in the ball  $B_{r_2}(u_2)$ , and continue this process until all  $v_i$ 's are deleted. Let  $t$  be the number of steps in this process. By the nature of this procedure, it is clear that  $r_1 > r_2 > \dots > r_t$ , and for  $i < j$  we have  $u_j \notin B_{r_i}(u_i)$ . Thus, we obtain  $\|u_i - u_j\| > r_i = \|u_i\| > \|u_j\|$  for every  $i > j$  which implies that 0 is the NN of every  $u_i$  for the set of points  $\{u_1, \dots, u_t, 0\}$ . Therefore, we obtain  $t \leq \kappa'$ . Moreover, note that we delete at most  $k$  of the  $v_i$ 's at each step. Because, 0 is a  $k$ NN to each  $u_i$  and there can be at most  $k - 1$  points in the ball  $B_{r_i}(u_i)$  other than  $u_i$  itself, since 0 lies on the boundary of the ball. Finally, we have the desired result  $m \leq \kappa'k$ .  $\square$

**Lemma 2.3.3.** *For every  $n \geq k(\kappa' + 1) + 2$ , we have  $P(Q_j^{[k]}(\mathcal{U}_n) > 0)$  and  $P(Q_j^{[k]}(\mathcal{P}_n) > 0)$  are positive if and only if  $j \leq \kappa'k$ .*

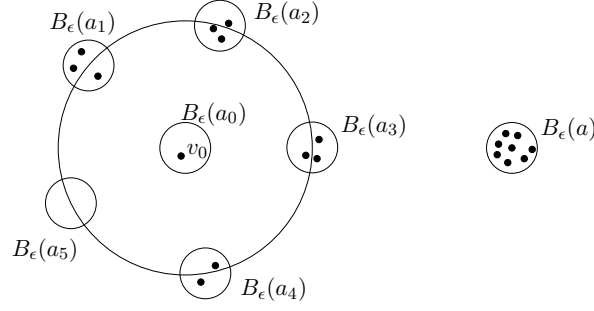


Figure 2.1: An illustration of  $n$  points with  $d_{in}(v_0) = j$  for  $d = 2$ ,  $n = 20$ ,  $k = 3$  and  $j = 11$ . Notice that  $s_1 = s_2 = s_3 = 3$ ,  $s_4 = 2$  and  $s_5 = 0$ .

*Proof.* As an immediate consequence of Lemma 2.3.2, we see that  $Q_j^{[k]}$  is identically zero for every  $j > \kappa'k$ .

For the values of  $j$  with  $j \leq \kappa'k$ , recall that there exist points  $a_1, a_2, \dots, a_{\kappa'}$  lying on the boundary of the unit sphere in  $\mathbb{R}^d$  such that  $\|a_i - a_j\| > 1$  for all  $1 \leq i \neq j \leq \kappa'$  by the definition of  $\kappa'$ . Let  $r = \min\{\|a_i - a_j\| : 1 \leq i \neq j \leq \kappa'\}$  and  $\epsilon$  be a positive number less than  $\min\{(r-1)/4, 1/4\}$ . Let  $a_0 = 0$  and  $a = (3, 0, 0, \dots, 0) \in \mathbb{R}^d$ .

Let  $V$  be a set of  $n$  points such that  $\text{card}(V \cap B_\epsilon(a_0)) = 1$ ,  $\text{card}(V \cap B_\epsilon(a_i)) = s_i \leq k$  for every  $1 \leq i \leq \kappa'$  and all the remaining points of  $V$  are in  $B_\epsilon(a)$ . See Figure 2.1 for an example. Let  $v_i \in B_\epsilon(a_i)$  for every  $i$  and  $v \in B_\epsilon(a)$ . Then we have

$$\|v_i - v_j\| > \|a_i - a_j\| - 2\epsilon \geq r - 2\epsilon > 1 + 2\epsilon > \|v_i - v_0\| \quad (2.11)$$

and

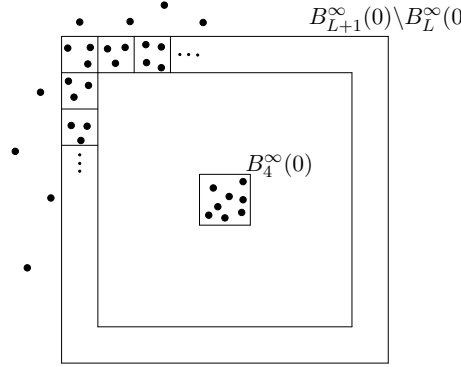
$$\|v_i - v\| > \|a_i - a\| - 2\epsilon \geq 2 - 2\epsilon > 1 + 2\epsilon > \|v_i - v_0\| \quad (2.12)$$

for all  $1 \leq i \neq j \leq \kappa'$ . In words, 0 is closer to  $v_i$  than the points in  $B_\epsilon(v_j)$  with  $j \neq i$  and  $B_\epsilon(a)$ . Therefore, since  $s_i \leq k$  for every  $i$ , results in (2.11) and (2.12) imply that 0 is a  $k$ NN to each  $v_i$  with  $i \geq 1$ . Also, for  $v, w \in V \cap B_\epsilon(a)$  we have

$$\|v - w\| < 2\epsilon < 3 - 2\epsilon < \|v - v_0\|,$$

which implies that  $w$  is closer to  $v$  than  $v_0$ . As  $n - (s_1 + \dots + s_{\kappa'}) - 1 \geq n - \kappa'k - 1 \geq k + 1$ ,  $V$  contains at least  $k + 1$  points in  $B_\epsilon(a)$ , and hence  $v_0$  is a  $k$ NN to none of the points in  $B_\epsilon(a)$ . Thus,  $v_0$  is a  $k$ NN to exactly  $j = s_1 + \dots + s_{\kappa'}$  points in  $V$ , that is,  $d_{in}(v_0) = j$  in  $D_{nn}^{[k]}(V)$ . Note that  $j$  can attain any value through 0 to  $\kappa'k$  for convenient values of  $s_i$ 's.

Clearly, having a scaled and translated version of such a configuration described above under  $\mathcal{P}_n$  or  $\mathcal{U}_n$  is of positive probability, and therefore, desired result follows.  $\square$

Figure 2.2: An illustration of the event  $E_1 \cap E_2$  for  $d = 2$  and  $k = 2$ .

### 2.3.2 Asymptotic Distribution of $Q_j^{[k]}$

We now obtain LLN and CLT results for  $Q_j^{[k]}$  for every  $0 \leq j \leq \kappa'k$  by Theorem 2.1.1.

**Corollary 2.3.4** (LLN and CLT for  $Q_j^{[k]}$ ). *For every  $0 \leq j \leq \kappa'k$ , there exist constants  $q_j(d, k)$ ,  $\tau_j^2 := \tau_j^2(d, k)$  and  $\sigma_j^2 := \sigma_j^2(d, k)$  with  $0 < \tau_j^2 \leq \sigma_j^2$  such that as  $n \rightarrow \infty$ ,*

$$\begin{aligned} n^{-1}Q_j^{[k]}(\mathcal{U}_n) &\xrightarrow{c.m.c.c.} q_j(d, k), \\ n^{-1}\mathbf{Var}(Q_j^{[k]}(\mathcal{U}_n)) &\rightarrow \tau_j^2, \\ n^{-1/2}(Q_j^{[k]}(\mathcal{U}_n) - \mathbf{E}(Q_j^{[k]}(\mathcal{U}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau_j^2), \\ n^{-1}Q_j^{[k]}(\mathcal{P}_n) &\xrightarrow{c.m.c.c.} q_j(d, k), \\ n^{-1}\mathbf{Var}(Q_j^{[k]}(\mathcal{P}_n)) &\rightarrow \sigma_j^2, \\ n^{-1/2}(Q_j^{[k]}(\mathcal{P}_n) - \mathbf{E}(Q_j^{[k]}(\mathcal{P}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_j^2). \end{aligned}$$

Moreover,  $q_j(d, k)$ ,  $\tau_j^2$  and  $\sigma_j^2$  are independent of the choice of the support  $B_0$ .

*Proof.* Let  $D_i$  be the digraph with  $V(D_i) = \{1, 2, \dots, i+1\}$  and  $A(D_i) = \{(1, i+1), (2, i+1), \dots, (i, i+1)\}$ , for every  $1 \leq i \leq \kappa'(d)k$ . Also, let  $D_0$  be the digraph with only one vertex, i.e.,  $V(D_0) = \{1\}$  and  $A(D_0) = \emptyset$ . Recall that by (2.1) we have

$$Q_j^{(k)}(V) = \sum_{i=j}^{\kappa'(d)k} (-1)^{i-j} \binom{i}{j} H_{D_i}(V), \quad (2.13)$$

for every  $0 \leq j \leq \kappa'(d)k$  and finite point set  $V$ . Then the desired asymptotic results follow by Theorem 2.1.1 and (2.13).

To show that both  $\tau_j^2$  and  $\sigma_j^2$  are positive, it suffices to prove that  $\Delta_{Q_j^{[k]}}(\infty)$  is non-degenerate for every  $0 \leq j \leq \kappa'k$  by Proposition 2.2.1. The main idea in the proof is to present two configurations with different add one costs for  $Q_j^{[k]}$ . In the configurations

we provide, the points near the origin are separated away from the other points, and so, the insertion of 0 changes only the  $k$ NN relations between the points around the origin.

Let  $L$  be a large positive integer. Let  $B_r^\infty(x)$  denote the corresponding  $l_\infty$  ball, that is  $B_r^\infty(x) = [-r, r]^d + x$ . Partition the annulus  $B_{L+1}^\infty(0) \setminus B_L^\infty(0)$  into a finite collection of unit cubes. Let  $E_1$  be the event that each unit cube in the partition contains at least  $k+1$  points from  $\mathcal{P}$  and  $E_2$  be the event that there is no point of  $\mathcal{P}$  in  $B_L^\infty(0) \setminus B_4^\infty(0)$ . Note that whenever both  $E_1$  and  $E_2$  occur and  $B_4^\infty(0)$  contains at least  $k+1$  points, any  $k$ NN of a point in  $B_4^\infty(0)$  lies in  $B_4^\infty(0)$  and every  $k$ NN of a point in  $\mathbb{R}^d \setminus B_L^\infty(0)$  lies in  $\mathbb{R}^d \setminus B_L^\infty(0)$  for large  $L$ . Therefore, in this case, the insertion of a point at the origin can only affect the  $k$ NN digraph of the points in  $B_4^\infty(0)$ . See Figure 2.2 for an illustration.

Let  $a_1, \dots, a_{\kappa'}$  and  $r$  be as defined in the proof of Lemma 2.3.3. Let  $b_i = (1+r)a_i/2$  for all  $1 \leq i \leq \kappa'$  and  $\epsilon$  be a positive real number less than  $(r-1)/8$ . Now fix  $j$  and let  $s_1, s_2, \dots, s_{\kappa'}$  be nonnegative integers not greater than  $k$  such that  $s_1 + \dots + s_{\kappa'} = j$ . Let  $E_3$  be the event that  $B_\epsilon(a_i)$  contains  $s_i$  points and  $B_\epsilon(b_i)$  contains  $k+1-s_i$  points for every  $1 \leq i \leq \kappa'$  and there are no other points in  $B_4^\infty(0)$ . See Figure 2.3 for an illustration.

When  $E_3$  occurs, it is easy to verify that  $k$ NNs of a point in  $B_\epsilon(a_i) \cup B_\epsilon(b_i)$  lies in the same union. Therefore, all the indegrees of the points in  $B_4^\infty(0)$  are  $k$ . Once the origin is inserted to the set, 0 becomes a  $k$ NN to the points in  $B_\epsilon(a_i)$  for each  $i$  and not a  $k$ NN to any point in  $B_\epsilon(b_i)$ . Thus, the indegree of 0 is  $s_1 + \dots + s_{\kappa'} = j$ . Then for  $j \neq k$ , we see that whenever  $E_1 \cap E_2 \cap E_3$  occurs,  $Q_j$  definitely increases. As the event  $E_1 \cap E_2 \cap E_3$  has a positive probability, we obtain

$$P(\Delta_{Q_j^{[k]}}(\infty) \geq 1) > 0 \text{ for every } j \neq k. \quad (2.14)$$

Now let  $E_4$  be the event that there are  $k$  points in  $B_\epsilon(a_1)$ , one point in  $B_\epsilon(b_1)$  and no other points in  $B_4^\infty(0)$ . After the addition of 0, the indegree of the point in  $B_\epsilon(b_1)$  becomes 0, the indegree of each of the points in  $B_\epsilon(a_1)$  increases to  $k+1$  and the indegree of 0 is  $k$ . Since  $E_1 \cap E_2 \cap E_4$  is an event with positive probability, we have

$$P(\Delta_{Q_0^{[k]}}(\infty) = 1) > 0, P(\Delta_{Q_k^{[k]}}(\infty) = -k) > 0, P(\Delta_{Q_{k+1}^{[k]}}(\infty) = k) > 0, \quad (2.15)$$

and

$$P(\Delta_{Q_j^{[k]}}(\infty) = 0) > 0 \text{ for each } j \notin \{0, k, k+1\}. \quad (2.16)$$

Then by the results in (2.14) and (2.16), we obtain that  $\Delta_{Q_j^{[k]}}$  is non-degenerate for every  $0 \leq j \leq \kappa'$  with  $j \notin \{0, k, k+1\}$ .

Next let  $E_5$  be the event that each of  $B_\epsilon(-0.4a_1)$ ,  $B_\epsilon(0.6a_1)$  and  $B_\epsilon(1.3a_1)$  has one point,  $B_\epsilon(0.5a_1)$  contains  $k-1$  points and there is no other point in  $B_4^\infty(0)$ , where

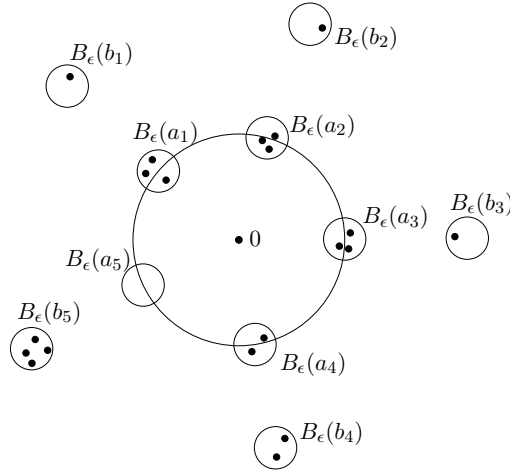


Figure 2.3: An illustration of the event  $E_3$  for  $d = 2$ ,  $k = 3$ ,  $j = 11$ ,  $s_1 = s_2 = s_3 = 3$ ,  $s_4 = 2$  and  $s_5 = 0$ .

$0 < \epsilon < 0.1$ . See Figure 2.4 for an example. One can easily see that the indegrees of the points in  $B_\epsilon(-0.4a_1)$ ,  $B_\epsilon(0.6a_1)$  and  $B_\epsilon(1.3a_1)$  are  $0$ ,  $k+1$  and  $k$ , respectively, and every point in  $B_\epsilon(0.5a_1)$  is of indegree  $k+1$ . After the addition of the origin, the indegrees of the points in  $B_\epsilon(-0.4a_1)$ ,  $B_\epsilon(0.6a_1)$  and  $B_\epsilon(1.3a_1)$  becomes  $1$ ,  $k$  and  $0$ , respectively, and the indegree of every point in  $B_\epsilon(0.5a_1)$  increases to  $k+2$ . Also,  $0$  is of indegree  $k+1$ . Therefore, as  $E_1 \cap E_2 \cap E_5$  is an event with positive probability, we have

$$P(\Delta_{Q_0^{[k]}}(\infty) = 0) > 0, P(\Delta_{Q_k^{[k]}}(\infty) \geq 0) > 0 \text{ and } P(\Delta_{Q_{k+1}^{[k]}}(\infty) = -(k-1)) > 0. \quad (2.17)$$

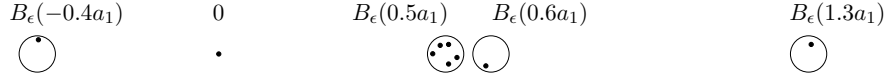
Then the results in (2.15) and (2.17) imply that  $\Delta_{Q_j^{[k]}}$  is non-degenerate as well whenever  $j \in \{0, k, k+1\}$ , and desired result follows.  $\square$

Computing the exact values of the constants  $q_j(d, k)$ ,  $\tau_j^2$  and  $\sigma_j^2$  analytically is tedious, if possible at all. For the case  $k = 1$ , the results in Newman et al. (1983) and Henze (1987) imply

$$q_j(d, 1) = \frac{1}{j!} \sum_{i=0}^{\kappa'-j} \frac{1}{i!} (-1)^i b_{j+i}(d)$$

for every  $0 \leq j \leq \kappa'$ , where  $b_0(d) = b_1(d) = 1$ ,

$$b_s(d) = \int \cdots \int_{\Gamma_s} \exp \left[ -m \left( \bigcup_{i=1}^s B_{\|x_i\|}(x_i) \right) \right] dx_1 \cdots dx_s,$$

Figure 2.4: An illustration of the event  $E_5$  for  $d = 2$  and  $k = 6$ .

and

$$\Gamma_s = \left\{ (x_1, \dots, x_s) : x_i \in \mathbb{R}^d, \|x_i\| < \min_{1 \leq l \leq s, l \neq i} \|x_i - x_l\|, 1 \leq i \leq s \right\}$$

for each  $2 \leq s \leq \kappa'$ . The values of  $b_s(d)$  are only approximated by Monte Carlo simulations. For the two dimensional case, Cuzick and Edwards (1990) provide

$$\begin{aligned} q_0(2, 1) &\approx 0.284, \quad q_1(2, 1) \approx 0.463, \quad q_2(2, 1) \approx 0.221, \\ q_3(2, 1) &\approx 3.04 \times 10^{-2}, \quad q_4(2, 1) \approx 6.58 \times 10^{-4}, \quad q_5(2, 1) \approx 1.90 \times 10^{-7}. \end{aligned}$$

On the other hand, for  $d = 1$  and  $k = 1$  all the limit values are known. In Bahadır and Ceyhan (2016) we have  $q_0(1, 1) = q_2(1, 1) = 1/4$ ,  $q_1(1, 1) = 1/2$ ,  $\tau_0^2(1, 1) = \tau_2^2(1, 1) = 19/240$  and  $\tau_1^2(1, 1) = 19/60$ . Using the results in Bahadır and Ceyhan (2016), one can easily show that  $\sigma_0^2(1, 1) = \sigma_2^2(1, 1) = 17/120$  and  $\sigma_1^2(1, 1) = 17/30$ .

Moreover, limiting value of  $q_j(d, k)$  as  $d \rightarrow \infty$  is studied by some authors. Newman et al. (1983) focus on the case  $k = 1$  and show that

$$\lim_{d \rightarrow \infty} q_j(d, 1) = \frac{e^{-1}}{j!}$$

for every  $j \geq 0$ , whereas Yao and Simons (1996) provide the answer for all  $k$ , that is,

$$\lim_{d \rightarrow \infty} q_j(d, k) = \frac{e^{-k} k^j}{j!}$$

for every  $j \geq 0$ .

### 2.3.3 Joint Distribution of $Q_j^{[k]}$ 's

Let  $Q_{in}^{[k]} = (Q_0^{[k]}, \dots, Q_{\kappa'k}^{[k]})$  and  $\Sigma_{Q_{in}^{[k]}}(\mathcal{U}_n)$  (resp.  $\Sigma_{Q_{in}^{[k]}}(\mathcal{P}_n)$ ) be the covariance matrix of  $Q_{in}^{[k]}(\mathcal{U}_n)$  (resp.  $Q_{in}^{[k]}(\mathcal{P}_n)$ ). Note that for any two random variables  $X$  and  $Y$ , we have  $\mathbf{Cov}(X, Y) = (\mathbf{Var}(X+Y) - \mathbf{Var}(X-Y))/4$ . Since any linear combination of  $Q_j^{[k]}$ 's is a linear combination of  $H_{D_i}$ 's defined in the proof of Corollary 2.3.4, Theorem 2.1.1 implies that  $\mathbf{Cov}(Q_{j_1}^{[k]}, Q_{j_2}^{[k]})/n$  converges to a constant as  $n \rightarrow \infty$  for any  $0 \leq j_1, j_2 \leq \kappa'k$ . Therefore, there exist constant  $(1 + \kappa'k) \times (1 + \kappa'k)$  matrices  $\Sigma_{\mathcal{U}} := \Sigma_{\mathcal{U}}(k, d)$  and

$\Sigma_{\mathcal{P}} := \Sigma_{\mathcal{P}}(k, d)$  such that

$$\Sigma_{Q_{in}^{[k]}(\mathcal{U}_n)/n} \rightarrow \Sigma_{\mathcal{U}} \text{ and } \Sigma_{Q_{in}^{[k]}(\mathcal{P}_n)/n} \rightarrow \Sigma_{\mathcal{P}}, \quad (2.18)$$

as  $n \rightarrow \infty$ . Similarly, we see that any linear combination of  $Q_j^{[k]}/\sqrt{n}$ 's converges in law to a normal variable, and therefore, Cramer-Wold device yields the following corollary.

**Corollary 2.3.5.** *Let  $\Sigma_{\mathcal{U}}$  and  $\Sigma_{\mathcal{P}}$  be the limiting matrices in (2.18). Then, as  $n \rightarrow \infty$  we have*

$$n^{-1/2}(Q_{in}^{[k]}(\mathcal{U}_n) - \mathbf{E}(Q_{in}^{[k]}(\mathcal{U}_n))) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma_{\mathcal{U}})$$

and

$$n^{-1/2}(Q_{in}^{[k]}(\mathcal{P}_n) - \mathbf{E}(Q_{in}^{[k]}(\mathcal{P}_n))) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma_{\mathcal{P}}),$$

where  $\mathbf{0}$  is the zero vector in  $\mathbb{R}^{1+\kappa'k}$  and  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  stands for the multivariate normal variable with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ .

Next, we study the ranks of the covariance matrices  $\Sigma_{Q_{in}^{[k]}(\mathcal{U}_n)}$  and  $\Sigma_{Q_{in}^{[k]}(\mathcal{P}_n)}$ . Since  $Q_j^{[k]}(V)$  counts the number of vertices in  $D_{nn}^{[k]}(V)$  with indegree  $j$ , we have

$$\text{card}(V) = \sum_{j=0}^{\kappa'k} Q_j^{[k]}(V). \quad (2.19)$$

As the number of arcs is equal to the sum of all the indegrees in a digraph, considering the  $D_{nn}^{[k]}(V)$  gives

$$k \cdot \text{card}(V) = \sum_{j=0}^{\kappa'k} j Q_j^{[k]}(V). \quad (2.20)$$

The equations in (2.19) and (2.20) appears to be the only linear relations between  $Q_j^{[k]}(V)$ 's and  $\text{card}(V)$ . Combining the results in these equations yields

$$\sum_{j=0}^{\kappa'k} (j - k) Q_j^{[k]} = 0. \quad (2.21)$$

Note that Equation (2.21) provides a non-trivial linear dependence relation for  $Q_j^{[k]}$ 's.

Now, suppose that  $\Sigma_{Q_{in}^{[k]}(\mathcal{X}_n)} \mathbf{a} = 0$  for some  $(1 + \kappa'k) \times 1$  real vector  $\mathbf{a}$ , where  $\mathcal{X}_n$  is

$\mathcal{U}_n$  or  $\mathcal{P}_n$ . Then, note that

$$0 = \mathbf{a}^t \Sigma_{Q_{in}^{[k]}(\mathcal{X}_n)} \mathbf{a} = \mathbf{Var} \left( \sum_{j=0}^{\kappa'k} a_j Q_j^{[k]}(\mathcal{X}_n) \right),$$

where  $\mathbf{a}^t = (a_0, \dots, a_{\kappa'k})$  is the transpose of the vector  $\mathbf{a}$ . So, we obtain that  $\sum_{j=0}^{\kappa'k} a_j Q_j^{[k]}(\mathcal{X}_n)$  is non-random as its variance is 0. Notice that letting  $\mathbf{a}^t$  to be any scalar multiple of  $(k, k-1, \dots, 1, 0, -1, \dots, (k-\kappa'k))$  satisfies the assumption for  $\mathbf{a}$  because of Equation (2.21).

Recall that  $\text{card}(\mathcal{U}_n) = n$  which is non-random, and therefore we can also take  $\mathbf{a}^t = (1, 1, \dots, 1)$  by (2.19). Thus, the rank of  $\Sigma_{Q_{in}^{[k]}(\mathcal{U}_n)}$  is at most  $(1 + \kappa'k) - 2 = \kappa'k - 1$ . But, on the other hand,  $\text{card}(\mathcal{P}_n)$  has a Poisson distribution with mean and variance equal to  $n$ . Hence, Equation (2.19) or (2.20) does not provide another example for  $\mathbf{a}$ , and so we can only state that the rank of  $\Sigma_{Q_{in}^{[k]}(\mathcal{P}_n)}$  is at most  $(1 + \kappa'k) - 1 = \kappa'k$ . We strongly believe that the upper bounds we provided for the rank of the covariance matrices are actually equalities, for every  $n \geq k(\kappa' + 1) + 2$ . Furthermore, the limiting matrices  $\Sigma_{\mathcal{U}}$  and  $\Sigma_{\mathcal{P}}$  seem to have the same corresponding ranks. Yet, these assertions currently remain as conjectures.

## 2.4 Number of Shared $k$ NNs

In this section we study the asymptotic distribution of  $Q^{[k]}$ . Recall that  $Q^{[k]} = H_D$  where  $D$  is the digraph with vertex set  $\{1, 2, 3\}$  and arc set  $\{(1, 2), (3, 2)\}$ . Thus, we can apply Theorem 2.1.1 for  $Q^{[k]}$ . Moreover, the events  $E_1 \cap E_2 \cap E_4$  and  $E_1 \cap E_2 \cap E_5$  described in the proof of Corollary 2.3.4 give

$$P(\Delta_{Q^{[k]}}(\infty) = k^2) > 0 \text{ and } P(\Delta_{Q^{[k]}}(\infty) = k^2 - 1) > 0,$$

respectively, and hence, we get  $\Delta_{Q^{[k]}}(\infty)$  is non-degenerate. So, by Proposition 2.2.1, the limiting variance values are positive, and we obtain the following result.

**Corollary 2.4.1** (LLN and CLT for  $Q^{[k]}$ ). *There exist constants  $q(d, k), \tau_Q^2 := \tau_Q^2(d, k)$  and  $\sigma_Q^2 := \sigma_Q^2(d, k)$  with  $0 < \tau_Q^2 \leq \sigma_Q^2$  such that as  $n \rightarrow \infty$ ,*

$$\begin{aligned} n^{-1} Q^{[k]}(\mathcal{U}_n) &\xrightarrow{\text{c.m.c.c.}} q(d, k), \\ n^{-1} \mathbf{Var}(Q^{[k]}(\mathcal{U}_n)) &\rightarrow \tau_Q^2, \\ n^{-1/2} (Q^{[k]}(\mathcal{U}_n) - \mathbf{E}(Q^{[k]}(\mathcal{U}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau_Q^2), \\ n^{-1} Q^{[k]}(\mathcal{P}_n) &\xrightarrow{\text{c.m.c.c.}} q(d, k), \\ n^{-1} \mathbf{Var}(Q^{[k]}(\mathcal{P}_n)) &\rightarrow \sigma_Q^2, \\ n^{-1/2} (Q^{[k]}(\mathcal{P}_n) - \mathbf{E}(Q^{[k]}(\mathcal{P}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_Q^2). \end{aligned}$$

Table 2.1: Values of  $q(d, k)$  for  $d = 1, 2$  and  $k = 1, \dots, 5$ .

| $d$ | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ |
|-----|---------|---------|---------|---------|---------|
| 1   | 0.25    | 1.5     | 3.75    | 7       | 11.25   |
| 2   | 0.3166  | 1.5868  | 3.8484  | 7.1079  | 11.3667 |

Additionally,  $q(d, k)$ ,  $\tau_Q^2$  and  $\sigma_Q^2$  are independent of the choice of the support  $B_0$ .

For  $d = 1$ , Theorem 3.2 in Schilling (1986b) implies  $q(1, k) = k^2/2 - k/4$  for each  $k$ . By Bahadır and Ceyhan (2016), we also have  $\tau_Q^2(1, 1) = 19/240$  and  $\sigma_Q^2(1, 1) = 17/120$ .

For  $d \geq 2$ , we only have numerical approximations for  $q(d, k)$ . For example, Cuzick and Edwards (1990) provides

$$q(2, 1) \approx 0.3166, q(2, 2) \approx 1.58685, q(2, 3) \approx 3.84845. \quad (2.22)$$

On the other hand, the results of Schilling (1986b) give

$$\begin{aligned} q(2, 1) &\approx 0.315, q(2, 2) \approx 1.575, q(2, 3) \approx 3.82, \\ q(3, 1) &\approx 0.355, q(3, 2) \approx 1.645, q(3, 3) \approx 3.93. \end{aligned} \quad (2.23)$$

Notice that the results in (2.22) and (2.23) differ slightly. Monte Carlo estimations we perform for  $q(2, k)$  for  $k = 1, 2, 3$  are closer to the ones in (2.22). For  $k = 1, \dots, 5$ , exact value of  $q(1, k)$  and the value of  $q(2, k)$  obtained in Cuzick and Edwards (1990) are presented in Table 2.1.

Furthermore, results in Schilling (1986b) implies

$$\lim_{d \rightarrow \infty} q(d, k) = \frac{k^2}{2},$$

for any  $k$ . Let  $V = \{v_1, \dots, v_s\}$  and  $d_i$  denote the indegree of  $v_i$  in  $D_{nn}^{[k]}(V)$ . Then, it is easy to see that the number of shared  $k$ NNs is  $\sum_{i=1}^s d_i(d_i - 1)/2$ . Moreover, a double counting argument for the number of arcs in the  $k$ NND gives  $\sum_{i=1}^s d_i = sk$ . Thus, we get  $\sum_{i=1}^s d_i(d_i - 1)/2 = s(k^2 - k)/2 + \sum_{i=1}^s (d_i - k)^2/2 \geq s(k^2 - k)/2$  which yields

$$Q^{[k]}(V) \geq \frac{\text{card}(V) \cdot (k^2 - k)}{2}, \quad (2.24)$$

for any finite point set  $V$ . By (2.24), one can easily obtain

$$q(d, k) \geq \frac{k(k - 1)}{2}, \quad (2.25)$$

for every  $d$  and  $k$ .

Recall that

$$Q^{[k]} = \sum_{j=0}^{\kappa'k} \binom{j}{2} Q_j^{[k]},$$

and so, we have

$$q(d, k) = \sum_{j=0}^{\kappa'k} \binom{j}{2} q_j(d, k).$$

Then, it is easy to verify that  $q(d, 1) = b_2(d)/2$  where  $b_2(d)$  is as defined in Section 2.3.2. Yet, for  $d \geq 2$ , we only have approximation of  $b_2(d)$  based on Monte Carlo simulations.

## 2.5 Number of Reflexive $k$ NNs

In this section, we study the asymptotic behavior of number of reflexive pairs  $R^{[k]}$ .

**Corollary 2.5.1** (LLN and CLT for  $R^{[k]}$ ). *There exist constants  $r(d, k), \tau_R^2 := \tau_R^2(d, k)$  and  $\sigma_R^2 := \sigma_R^2(d, k)$  with  $0 < \tau_R^2 \leq \sigma_R^2$  such that as  $n \rightarrow \infty$ ,*

$$\begin{aligned} n^{-1} R^{[k]}(\mathcal{U}_n) &\xrightarrow{c.m.c.c.} r(d, k), \\ n^{-1} \mathbf{Var}(R^{[k]}(\mathcal{U}_n)) &\rightarrow \tau_R^2, \\ n^{-1/2}(R^{[k]}(\mathcal{U}_n) - \mathbf{E}(R^{[k]}(\mathcal{U}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau_R^2), \\ n^{-1} R^{[k]}(\mathcal{P}_n) &\xrightarrow{c.m.c.c.} r(d, k), \\ n^{-1} \mathbf{Var}(R^{[k]}(\mathcal{P}_n)) &\rightarrow \sigma_R^2, \\ n^{-1/2}(R^{[k]}(\mathcal{P}_n) - \mathbf{E}(R^{[k]}(\mathcal{P}_n))) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_R^2). \end{aligned}$$

Furthermore,  $r(d, k), \tau_R^2$  and  $\sigma_R^2$  are independent of the choice of the support  $B_0$ .

*Proof.* Recall that  $R^{[k]} = H_D$  where  $D$  is the digraph with vertex set  $\{1, 2\}$  and arc set  $\{(1, 2), (2, 1)\}$ , and therefore, Theorem 2.1.1 yields the asymptotic results.

Furthermore, we also show that  $\Delta_R(\infty)$  is non-degenerate which implies  $0 < \tau_R^2$  and hence  $0 < \sigma_R^2$ . Both of the events  $E_1 \cap E_2 \cap E_4$  and  $E_1 \cap E_2 \cap E_5$  defined in the proof of Corollary 2.3.4 yield

$$P(\Delta_{R^{[k]}}(\infty) = 0) > 0. \quad (2.26)$$

Letting  $s_1 = s_2 = k$  and  $s_i = 0$  for all  $i \geq 3$  for the event  $E_3$  in the proof of Corollary 2.3.4 gives

$$P(\Delta_{R^{[k]}}(\infty) = k) > 0. \quad (2.27)$$

Table 2.2: Values of  $r(d, k)$  for  $d = 1, 2$  and  $k = 1, \dots, 5$ .

| $d$ | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ |
|-----|---------|---------|---------|---------|---------|
| 1   | 0.3333  | 0.7407  | 1.1728  | 1.6168  | 2.0680  |
| 2   | 0.3107  | 0.7105  | 1.1365  | 1.5751  | 2.0215  |

Then by (2.26) and (2.27), we see that  $\Delta_{R[k]}(\infty)$  is not degenerate.  $\square$

By the results in Henze (1987), Pickard (1982) and Schilling (1986b), we obtain

$$r(d, k) = \sum_{s=1}^k \sum_{t=1}^k r_d(s, t),$$

where

$$r_d(s, t) = \frac{\omega(d)}{2} \sum_{i=0}^{\min\{s-1, t-1\}} \frac{(s+t-i-2)!}{i!(s-i-1)!(t-i-1)!} (2\omega(d)-1)^i (1-\omega(d))^{s+t-2i-2}$$

and

$$\omega(d) = \begin{cases} \left[ \frac{3}{2} + \frac{1}{2} \sum_{i=1}^m \frac{1 \cdot 3 \cdots (2i-1)}{2 \cdot 4 \cdots (2i)} \left( \frac{3}{4} \right)^i \right]^{-1} & \text{if } d = 2m+1, \\ \left[ \frac{4}{3} + \frac{\sqrt{3}}{2\pi} \left( 1 + \sum_{i=1}^{m-1} \frac{2 \cdot 4 \cdots (2i)}{3 \cdot 5 \cdots (2i+1)} \left( \frac{3}{4} \right)^i \right) \right]^{-1} & \text{if } d = 2m. \end{cases}$$

From a geometric point of view,  $\omega(d)$  is the volume of a unit sphere in  $\mathbb{R}^d$  divided by the volume of the union of two such spheres whose centres are separated by a distance 1. The rounded values of  $r(d, k)$  for  $d = 1, 2$  and  $k = 1, \dots, 5$  are presented in Table 2.2.

However, almost nothing about the exact values of limiting variances  $\tau_R^2$  and  $\sigma_R^2$  exists in literature. We only have  $\tau_R^2(1, 1) = 2/45$  and  $\sigma_R^2(1, 1) = 7/45$  by the results in Bahadır and Ceyhan (2016).

Finally, we present an interesting connection between reflexive pairs and components of a 1NN digraph which is also stated in Eppstein et al. (1997) and Enns et al. (1999).

**Proposition 2.5.2.** *Let  $V$  be a finite point set including at least two points such that pairwise distances between the points of  $V$  are all distinct. Then, each weakly connected component of 1NN digraph of  $V$  contains exactly one reflexive pair and hence, number of weakly connected components of  $D_{nn}^{[1]}(V)$  is  $R^{[1]}(V)$ .*

*Proof.* Construct the 1NN digraph of  $V$ . We first show that any weakly connected com-

ponent of  $D_{nn}^{[1]}(V)$  contains at least one reflexive pair. Note that any weakly connected component contains at least two points and therefore, it also contains at least one arc. Pick any component and the shortest arc in it. Here by *length* of an arc we refer to the distance between the endpoints of the arc, and the shortest arc is the one with the minimum length. As there are finitely many arcs in the component, the shortest arc exists, say  $(u, v)$ . Then, by definition we see that  $NN$  of  $u$  is  $v$ . Let  $NN$  of  $v$  be  $w$ . Clearly the arc  $(v, w)$  belongs to this component and hence, we get  $\|v - w\| \geq \|v - u\|$  by the choice of  $(u, v)$ . On the other hand, considering the definition of  $NN$  for  $v$  yields  $\|v - w\| \leq \|v - u\|$ . Thus, by the uniqueness of the  $NN$  of any point we obtain that  $w = u$  which implies that  $\{u, v\}$  is a reflexive pair in this component.

Next suppose that a component contains two reflexive pairs  $\{u, v\}$  and  $\{w, z\}$ . As each point has a unique  $NN$ , we have  $\{u, v\} \cap \{w, z\} = \emptyset$ . In the underlying graph of  $D_{nn}^{[1]}(V)$ , consider a shortest path with one end point from  $\{u, v\}$  and the other one from  $\{w, z\}$ , say  $v_1, \dots, v_s$ . Since the edges corresponding to the reflexive pairs appear in the same component, such paths exist and as the graph is finite, a shortest one is well defined. Without loss of generality, we may assume that  $v_1 = v$  and  $v_s = z$ . Note that by the choice of the path we have  $v_2 \neq u$ . Moreover, since we have the edge  $v_1 v_2$  in the underlying graph, we see that  $1NN$  of  $v_1$  is  $v_2$  or  $1NN$  of  $v_2$  is  $v_1$ . On the other hand, we have  $v_1 = v$  and  $1NN$  of  $v$  is  $u \neq v_2$ , and hence, we obtain that  $1NN$  of  $v_2$  is  $v_1$ . In the same manner, we see that one of  $v_2$  and  $v_3$  is the  $1NN$  of the other one, and therefore, we get  $1NN$  of  $v_3$  is  $v_2$  since  $1NN$  of  $v_2$  is  $v_1$  and  $v_1 \neq v_3$ . By induction on the indices of  $v_i$ 's we obtain that the  $1NN$  of  $v_s$  is  $v_{s-1}$ ; that is,  $v_{s-1}$  is the  $1NN$  of  $z$ . Then, the uniqueness of  $1NN$  implies  $w = v_{s-1}$  which contradicts the choice of the path. Thus, each component has exactly one reflexive pair and therefore, the result follows.  $\square$

## 2.6 The Case of $d = k = 1$

In this section, we study the quantities for  $k = 1$  in one-dimensional data.

Clearly, in one dimension, a point is  $NN$  to at most two other points and hence  $Q_j^{[1]} = 0$  for every  $j \geq 3$ . Double counting arguments for the number of vertices and the number of arcs give

$$\text{card}(V) = Q_0^{[1]}(V) + Q_1^{[1]}(V) + Q_2^{[1]}(V) \quad (2.28)$$

and

$$\text{card}(V) = 0 \cdot Q_0^{[1]}(V) + 1 \cdot Q_1^{[1]}(V) + 2 \cdot Q_2^{[1]}(V) \quad (2.29)$$

for any finite point set  $V \subset \mathbb{R}$ , respectively. On the other hand, we have

$$Q^{[1]}(V) = \sum_{j=0}^2 \binom{j}{2} Q_j^{[1]}(V) = Q_2^{[1]}(V) \quad (2.30)$$

for any point set  $V$ . Therefore, the results in (2.28)-(2.30) give

$$Q_0^{[1]}(V) = Q_2^{[1]}(V) = Q^{[1]}(V) \text{ and } Q_1^{[1]}(V) = \text{card}(V) - 2 \cdot Q^{[1]}(V)$$

for every  $V$ , and hence, it suffices to study only  $Q_j^{[1]}$  instead of  $Q_j^{[1]}$ 's.

Let  $\mathcal{U}_n = \{U_1, U_2, \dots, U_n\}$  be a random sample of size  $n$  from the uniform distribution  $U(0, 1)$ . On the real line, there is a nice ordering structure for the data which we exploit in our results. Let  $U_{(1)}, U_{(2)}, \dots, U_{(n)}$  be the order statistics of  $\{U_1, U_2, \dots, U_n\}$ . Denote the spacings between the order statistics as  $D_i := U_{(i+1)} - U_{(i)}$  for  $1 \leq i \leq n-1$  with  $D_0 := U_{(1)}$ .

**Lemma 2.6.1.** *For  $n \geq 3$*

$$R^{[1]}(\mathcal{U}_n) = \mathbf{1}_{\{D_1 < D_2\}} + \sum_{i=2}^{n-2} \mathbf{1}_{\{D_i < \min\{D_{i+1}, D_{i-1}\}\}} + \mathbf{1}_{\{D_{n-1} < D_{n-2}\}},$$

and for  $n \geq 4$

$$Q^{[1]}(\mathcal{U}_n) = \mathbf{1}_{\{D_2 < D_3\}} + \sum_{i=2}^{n-3} \mathbf{1}_{\{D_i < D_{i-1}, D_{i+1} < D_{i+2}\}} + \mathbf{1}_{\{D_{n-2} < D_{n-3}\}},$$

where  $\mathbf{1}_A$  is the indicator for the event  $A$ .

*Proof.* First observe that the NNs of  $U_{(1)}$  and  $U_{(n)}$  are always  $U_{(2)}$  and  $U_{(n-1)}$ , respectively. Therefore,  $\{U_{(1)}, U_{(2)}\}$  is a reflexive pair if and only if  $D_1 < D_2$  and, similarly,  $\{U_{(n-1)}, U_{(n)}\}$  is reflexive if and only if  $D_{n-1} < D_{n-2}$ . Also note that, for each  $2 \leq i \leq n-1$ , the NN of  $U_{(i)}$  is either  $U_{(i-1)}$  (if  $D_{i-1} < D_i$ ) or  $U_{(i+1)}$  (if  $D_i < D_{i-1}$ ). Thus, for  $2 \leq i, j \leq n-1$ , the pair  $\{U_{(i)}, U_{(j)}\}$  with  $i < j$  is reflexive if and only if  $j = i+1$  and  $D_i$  is less than both  $D_{i-1}$  and  $D_{i+1}$ . So, we obtain the first identity in the Lemma 2.6.1. For the representation of  $Q^{[1]}(\mathcal{U}_n)$ , in a similar manner, one can easily see that  $U_{(i)}$  and  $U_{(j)}$  ( $i < j$ ) have the same NN only if  $j = i+2$  and the common NN is  $U_{(i+1)}$ , and obtain the desired result.  $\square$

As the quantities  $R^{[1]}(\mathcal{U}_n)$  and  $Q^{[1]}(\mathcal{U}_n)$  depend on the ordering of the spacings, we focus on the distribution of the spacings. By elementary probability arguments (e.g., Jacobian density theorem) it follows that the joint density of the spacings  $(D_0, D_1, \dots, D_{n-1})$  is

$$f_S(d_0, d_1, \dots, d_{n-1}) = n! \mathbf{1}_{\{d_0 + d_1 + \dots + d_{n-1} < 1\}} \mathbf{1}_{\{\min\{d_0, d_1, \dots, d_{n-1}\} > 0\}} \quad (2.31)$$

with the understanding that  $\{d_0, d_1, \dots, d_{n-1} > 0\} = \{d_0 > 0, d_1 > 0, \dots, d_{n-1} > 0\}$ . By (2.31) it is clear that the spacings  $D_1, \dots, D_{n-1}$  are interchangeable and hence  $P(D_{\sigma(1)} < \dots < D_{\sigma(n-1)}) = P(D_1 < \dots < D_{n-1})$  for any permutation  $\sigma$  in  $P_{n-1}$ , where  $P_{n-1}$  is the permutation group on  $\{1, 2, \dots, n-1\}$ . In other words, every ordering of the spacings  $D_1, \dots, D_{n-1}$  is equally likely to occur.

Let  $\sigma$  be chosen uniformly at random from  $P_{n-1}$ . Define the events  $A_1 = \{\sigma(1) < \sigma(2)\}$ ,  $A_{n-1} = \{\sigma(n-1) < \sigma(n-2)\}$ ,  $A_i = \{\sigma(i) < \sigma(i-1), \sigma(i) < \sigma(i+1)\}$  for all  $2 \leq i \leq n-2$ , and the events  $B_1 = \{\sigma(2) < \sigma(3)\}$ ,  $B_{n-2} = \{\sigma(n-2) < \sigma(n-3)\}$  and  $B_i = \{\sigma(i) < \sigma(i-1), \sigma(i+1) < \sigma(i+2)\}$  for each  $2 \leq i \leq n-3$ . Then, by Lemma 2.6.1 and the interchangeability of the spacings we have

$$R^{[1]}(\mathcal{U}_n) \stackrel{d}{=} \sum_{i=1}^{n-1} \mathbf{1}_{A_i} \text{ and } Q^{[1]}(\mathcal{U}_n) \stackrel{d}{=} \sum_{i=1}^{n-2} \mathbf{1}_{B_i}, \quad (2.32)$$

where  $\stackrel{d}{=}$  denotes equality in distribution. Therefore, throughout of this section, we consider  $\sum_{i=1}^{n-1} \mathbf{1}_{A_i}$  and  $\sum_{i=1}^{n-2} \mathbf{1}_{B_i}$  for the probabilistic results for  $R^{[1]}(\mathcal{U}_n)$  and  $Q^{[1]}(\mathcal{U}_n)$ , respectively.

### 2.6.1 Means and Variances of $R^{[1]}$ and $Q^{[1]}$

Enns et al. (1999) call a weakly connected component of a digraph as *society* and examine the number of societies in a uniform data of size  $n$  in one dimension. By the simple observation that each society contains exactly one reflexive pair, they convert the problem into the number of reflexive pairs and focus on the ranking of the spacings. Considering the spacing with the largest length, they derive a recurrence relation and obtain  $\mathbf{E}(R^{[1]}(\mathcal{U}_n)) = n/3$  for  $n \geq 3$  and  $\mathbf{Var}(R^{[1]}(\mathcal{U}_n)) = 2n/45$  for  $n \geq 5$  by using generating functions. We verify their results by following the idea in Romik (2011).

We obtain the mean and variance of  $R^{[1]}(\mathcal{U}_n)$  by computing those of  $\sum_{i=1}^{n-1} \mathbf{1}_{A_i}$ . The random variable  $\sum_{i=1}^{n-1} \mathbf{1}_{A_i}$  is closely related to the length of the longest alternating subsequence in a random permutation (see, e.g., Romik (2011), Houdré and Restrepo (2010)). For a sequence of pairwise distinct real numbers  $x_1, \dots, x_n$ , a subsequence  $x_{i_1}, \dots, x_{i_k}$  with  $1 \leq i_1 < \dots < i_k \leq n$  is called *alternating* if it satisfies

$$x_{i_1} > x_{i_2} < x_{i_3} > \dots x_{i_k}.$$

Note that there may be more than one alternating subsequence with the maximal length. For instance, the sequence 6, 4, 1, 3, 5, 2 has seven longest alternating subsequences, particularly (6, 1, 3, 2), (6, 1, 5, 2), (6, 4, 5, 2), (6, 3, 5, 2), (4, 1, 3, 2), (4, 1, 5, 2) and (4, 3, 5, 2). Let the random variable  $L_n^{as}$  be the maximal length of an alternating subsequence of  $\tau(1), \dots, \tau(n)$ , where  $\tau$  is a uniformly random permutation from  $P_n$ .

Also, for  $2 \leq k \leq n-1$ ,  $x_k$  is called *local minimum* (resp. *local maximum*) if

$x_k < \min\{x_{k-1}, x_{k+1}\}$  (resp.  $x_k > \max\{x_{k-1}, x_{k+1}\}$ ) (Romik (2011)). Note that the sum  $\sum_{i=2}^{n-2} \mathbf{1}_{A_i}$  is the number of local minima in  $\sigma(1), \dots, \sigma(n-1)$ , where  $\sigma$  is a uniformly random permutation from  $P_{n-1}$ . Romik (2011) shows that  $L_n^{as}$  is equal to  $1 + \mathbf{1}_{\{\tau(1) > \tau(2)\}}$  plus the number of local minimums and local maximums in a random permutation, and provides  $\mathbf{E}(L_n^{as}) = 2n/3 + 1/6$  and  $\mathbf{Var}(L_n^{as}) = 8n/45 - 13/180$  (which are also computed in Stanley (2008) and Houdré and Restrepo (2010) in different ways). Notice that the number of local minima and number of local maxima differ by at most one, and hence the number of local minima is about half of  $L_n^{as}$ . Therefore, we have  $\mathbf{E}(R^{[1]}(\mathcal{U}_n))/n \rightarrow 1/3$  and  $\mathbf{Var}(R^{[1]}(\mathcal{U}_n))/n \rightarrow 2/45$  as  $n \rightarrow \infty$ . In fact, the limits  $1/3$  and  $2/45$  are actually attained for every  $n \geq 5$ .

**Theorem 2.6.2.** *For a random sample of size  $n$  from  $U(0, 1)$ , the mean and the variance of the number of reflexive pairs,  $R^{[1]}(\mathcal{U}_n)$ , is  $n/3$  (for  $n \geq 3$ ) and  $2n/45$  (for  $n \geq 5$ ), respectively.*

The proof of Theorem 2.6.2 is given in Appendix B.1. By using Theorem 2.6.2, one can easily obtain the following result.

**Corollary 2.6.3.** *For a random sample from HPP of intensity  $n$  on  $(0, 1)$ , the mean and the variance of the number of reflexive pairs,  $R^{[1]}(\mathcal{P}_n)$ , is  $n/3 + \epsilon_1(n)$  and  $7n/45 + \epsilon_2(n)$ , respectively, such that  $\epsilon_1(n) \rightarrow 0$  and  $\epsilon_2(n) \rightarrow 0$  as  $n \rightarrow \infty$ .*

The proof of Corollary 2.6.3 is given in Appendix B.2.

The means and variances of  $Q^{[1]}(\mathcal{U}_n)$  and  $Q^{[1]}(\mathcal{P}_n)$  can be derived in a similar manner.

**Theorem 2.6.4.** *For a random sample of size  $n$  from  $U(0, 1)$ , the mean and the variance of the number of shared NNs,  $Q^{[1]}(\mathcal{U}_n)$ , is  $n/4$  (for  $n \geq 4$ ) and  $19n/240$  (for  $n \geq 7$ ), respectively. Moreover, for a random sample from HPP of intensity  $n$  on  $(0, 1)$ , the mean and the variance of the number of shared NNs,  $Q^{[1]}(\mathcal{P}_n)$ , is  $n/4 + \delta_1(n)$  and  $17n/120 + \delta_2(n)$ , respectively, such that  $\delta_1(n) \rightarrow 0$  and  $\delta_2(n) \rightarrow 0$  as  $n \rightarrow \infty$ .*

The proof of Theorem 2.6.4 is given in Appendix B.3.

*Remark 2.6.5.* Our extensive Monte Carlo simulations strongly suggest (asymptotically) same variance value ( $2/45$  and  $19/240$ ) whenever the underlying process is a distribution on  $\mathbb{R}$  with an a.e. continuous density. But, we can not apply the method used in this section when the distribution of the sample is not uniform, since we lose the interchangeability of the spacings. Also, notice that the method we use is not applicable for higher dimensions or larger values of  $k$ . Because, when  $d > 1$  we lose the ordering structure, and when  $k > 1$ ,  $k$ NN relations depend also on sum of the consecutive spacings.

### 2.6.2 A Recurrence Relation for the Exact Distribution of $R^{[1]}(\mathcal{U}_n)$

Recall that for a sequence of pairwise distinct real numbers  $x_1, \dots, x_n$ , we say  $x_k$  is a local minimum if  $x_k$  is less than its neighbors (i.e.,  $x_{k-1}$  and  $x_{k+1}$ ) for  $2 \leq k \leq n-1$ .

Let us also consider  $x_1$  (resp.  $x_n$ ) as a local minimum if  $x_1 < x_2$  (resp.  $x_n < x_{n-1}$ ). Then, notice that  $\sum_{i=1}^{n-1} \mathbf{1}_{A_i}$  is exactly the number of local minima in a uniformly random permutation from  $P_{n-1}$ . Let  $p(n, k)$  denote  $P(R^{[1]}(\mathcal{U}_n) = k)$  and set  $p(1, 0) = p(2, 1) = 1$ . Also, let  $m(n, k)$  be the number of permutations in  $P_n$  with exactly  $k$  local minimums. Notice that  $p(n, k) = m(n-1, k)/(n-1)!$ . Since the term 1 in the sequence is always a local minimum and any two local minimums are not adjacent, we see that  $p(n, 0) = 0$  for  $n \geq 2$  and  $p(n, k) = 0$  whenever  $k > n/2$ .

Any permutation in  $P_n$  can be uniquely obtained by increasing each element of a permutation in  $P_{n-1}$  by one and inserting the element 1 in one of the possible  $n$  places. In this process, inserting the element 1 into the sequence does not effect the number of local minimums if it is placed next to a local minimum, and otherwise, increases the number of local minimums by one. Therefore, we obtain

$$m(n, k) = 2k \cdot m(n-1, k) + (n-2(k-1)) \cdot m(n-1, k-1),$$

since any two local minimums are not adjacent. Thus, as  $m(n-1, k) = p(n, k)(n-1)!$ , we have

$$p(n+1, k) = \frac{2k}{n} p(n, k) + \frac{n-2k+2}{n} p(n, k-1), \quad (2.33)$$

for every  $n \geq 2$ . Therefore, the exact pmf of  $R^{[1]}(\mathcal{U}_n)$  can be computed for any  $n \geq 3$  by using the recursion given in (2.33).

Enns et al. (1999) consider the index of the spacing with the largest length (i.e., index  $i$  such that  $D_i = \max\{D_1, D_2, \dots, D_{n-1}\}$ ) and derive the following recurrence relation

$$p(n, k) = \frac{2}{n-1} p(n-1, k) + \sum_{i=2}^{n-2} \sum_{j=1}^{k-1} \frac{p(i, j) p(n-i, k-j)}{n-1},$$

for every  $n \geq 4$ . Then, using generating functions, they obtain the relation in (2.33).

Let  $p^*(n, k) = P(R^{[1]}(\mathcal{P}_n) = k)$  for every non-negative integer  $k$  and  $n$ . Then, we have

$$p^*(n, k) = \sum_{m=0}^{\infty} P(\text{card}(\mathcal{P}_n) = m) P(R^{[1]}(\mathcal{U}_m) = k) = \sum_{m=0}^{\infty} p(m, k) e^{-n} \frac{n^m}{m!}. \quad (2.34)$$

Therefore, by (2.33) and (2.34) the pmf of  $R^{[1]}(\mathcal{P}_n)$  can be numerically approximated.

## 2.7 Pairwise Dependence of $R^{[1]}(\mathcal{U}_n)$ , $Q^{[1]}(\mathcal{U}_n)$ and $Q_j^{[1]}(\mathcal{U}_n)$

In this section, we study the case  $k = 1$  where we can show that  $R^{[1]}(\mathcal{U}_n)$ ,  $Q^{[1]}(\mathcal{U}_n)$  and  $Q_j^{[1]}(\mathcal{U}_n)$  are pairwise dependent. For simplicity in the notation, let  $R$ ,  $Q$  and  $Q_j$  denote  $R^{[1]}(\mathcal{U}_n)$ ,  $Q^{[1]}(\mathcal{U}_n)$  and  $Q_j^{[1]}(\mathcal{U}_n)$ , respectively.

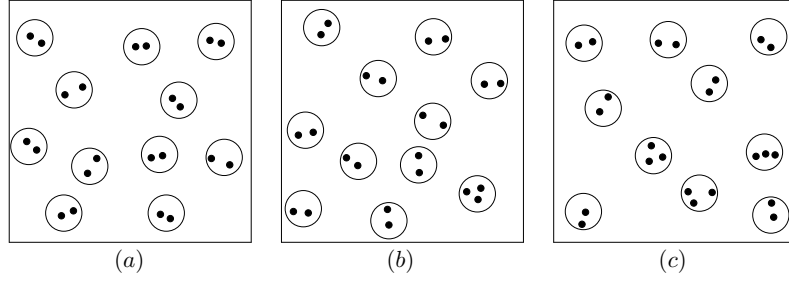


Figure 2.5: An illustration of  $n$  points with (a)  $n/2$  reflexive pairs for  $n = 22$ , (b)  $(n - 1)/2$  reflexive pairs for  $n = 23$ , and (c)  $(n - 3)/2$  reflexive pairs for  $n = 23$ .

We first show that  $R$  and  $Q_j$ 's are pairwise stochastically dependent for large  $n$ . We use the simple fact that if the random variables  $X$  and  $Y$  satisfy  $P(X \in A) > 0$ ,  $P(Y \in B) > 0$  and  $P(X \in A, Y \in B) = 0$  for some Borel sets  $A$  and  $B$ , then  $X$  and  $Y$  are dependent. Also recall that for  $n \geq 4$ , if the sample points  $x$  and  $y$  are sufficiently close to each other and far from the other  $n - 2$  sample points, then  $\{x, y\}$  forms a reflexive pair with each of indegree 1. In addition, for  $n \geq 5$ , if the sample points  $x, y, z$  are sufficiently close to each other and far from the remaining  $n - 3$  sample points, then the indegrees of  $x, y, z$  in the  $NND$  are 0, 1, 2 in some order.

**Theorem 2.7.1.** *For every  $n \geq \max\{9, \kappa' + 3\}$ ,  $R$  and  $Q_j$  are dependent for all  $0 \leq j \leq \kappa'$ . Moreover,  $R$  and  $Q$  are dependent for every  $n \geq 6$ .*

*Proof.* We first prove the dependence of  $R$  and  $Q_j$ 's. The proof is based on the parity of  $n$ . First assume that  $n$  is an even positive integer. It is easy to see that  $R = n/2$  if and only if sample points consist of  $n/2$  pairs which are pairwise far enough (i.e., members of each pair is NN to each other and each pair is sufficiently far from other such pairs). In this case, each point is of indegree 1. Therefore, we have

$$\{R = n/2\} \subseteq \{Q_1 = n\} \cap \left( \bigcap_{\substack{0 \leq j \leq \kappa' \\ j \neq 1}} \{Q_j = 0\} \right). \quad (2.35)$$

The event that each of  $n/2$  sufficiently small balls in the region contains exactly 2 sample points has a positive probability and therefore,  $P(R = n/2) > 0$  (See Figure 2.5 (a)). Moreover, by Lemma 2.3.3 we have  $P(Q_j \geq 1) > 0$  for every  $0 \leq j \leq \kappa'$ . By (2.35) we have  $\{R = n/2\} \cap \{Q_j \geq 1\} = \emptyset$  for each  $0 \leq j \neq 1 \leq \kappa'$ , and hence, we obtain that  $R$  and  $Q_j$  are dependent for every  $0 \leq j \neq 1 \leq \kappa'$ .

Next, note that as  $\sum_{j=0}^{\kappa'} Q_j = n$ , we have  $\{Q_0 > 0\} \subseteq \{Q_1 < n\}$ . We know that  $P(Q_0 > 0) > 0$  by Lemma 2.3.3, and therefore  $P(Q_1 < n) > 0$ . Moreover, the result in (2.35) gives  $\{R = n/2\} \cap \{Q_1 < n\} = \emptyset$ , and thus,  $R$  and  $Q_1$  are dependent as well.

Now suppose that  $n$  is an odd positive integer. The event that each of  $(n-3)/2$  sufficiently small balls in the region contains exactly 2 sample points and a sufficiently small ball contains exactly 3 sample points has a positive probability, and hence  $P(R = (n-1)/2) > 0$  (See Figure 2.5 (b)). Moreover, it is easy to verify that  $R = (n-1)/2$  implies

$$Q_0 = 1, Q_1 = n-2, Q_2 = 1 \text{ and } Q_j = 0 \text{ for all } 3 \leq j \leq \kappa'.$$

In other words, we have

$$\{R = (n-1)/2\} \subseteq \{Q_0 = Q_2 = 1\} \cap \{Q_1 = n-2\} \cap \left( \bigcap_{j=3}^{\kappa'} \{Q_j = 0\} \right). \quad (2.36)$$

So, by (2.36) we have  $\{R = (n-1)/2\} \cap \{Q_j \geq 1\} = \emptyset$  for all  $3 \leq j \leq \kappa'$ . Then since  $P(Q_j \geq 1) > 0$  by Lemma 2.3.3 and  $P(R = (n-1)/2) > 0$ , we obtain that  $R$  and  $Q_j$  are dependent for each  $3 \leq j \leq \kappa'$ .

Furthermore, the event that each of  $(n-9)/2$  sufficiently small balls in the region contains exactly 2 sample points and each of three sufficiently small balls contains exactly 3 sample points has a positive probability (See Figure 2.5 (c)), and in this case we have

$$R = (n-3)/2, Q_0 = 3, Q_1 = n-6 \text{ and } Q_2 = 3. \quad (2.37)$$

Therefore, each one of  $P(Q_0 = 3)$ ,  $P(Q_1 = n-6)$  and  $P(Q_2 = 3)$  is positive. By the results in (2.36) and (2.37), we have  $\{R = (n-1)/2\} \cap \{Q_0 = 3\} = \emptyset$ ,  $\{R = (n-1)/2\} \cap \{Q_1 = n-6\} = \emptyset$  and  $\{R = (n-1)/2\} \cap \{Q_2 = 3\} = \emptyset$ . Hence, we see that  $R$  and  $Q_j$  are dependent also for every  $0 \leq j \leq 2$ .

Finally, we prove that  $R$  and  $Q$  are dependent for each  $n \geq 6$ . Since  $Q = \sum_{j \geq 0} j(j-1)Q_j/2$ , by (2.35) we obtain

$$\{R = n/2\} \subseteq \{Q = 0\} \quad (2.38)$$

when  $n$  is even, and by (2.36) we have

$$\{R = (n-1)/2\} \subseteq \{Q = 1\} \quad (2.39)$$

when  $n$  is odd. Clearly, both  $R$  and  $Q$  always attain nonnegative integer values and  $R \leq n/2$  since each point has a unique  $NN$ . Hence, (2.38) and (2.39) imply

$$\{R \geq (n-1)/2\} \subseteq \{Q \leq 1\}. \quad (2.40)$$

On the other hand, consider the event that each of two sufficiently small balls contains three sample points and remaining  $n-6$  sample points are far enough from these two

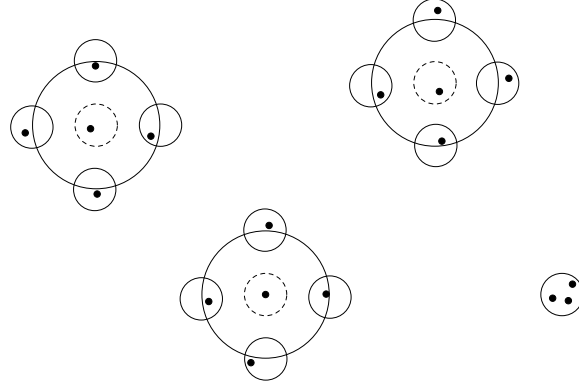


Figure 2.6: An illustration for  $d = 2$ ,  $n = 18$ ,  $j = 4$ ,  $t_4 = 3$  and  $s_4 = 3$ . Note that the sample points in the dashed circles are of indegree 4 and hence  $Q_4 \geq 3$ .

balls. In this case, the indegrees of the points in each one of the small balls are 0, 1 and 2, and thus,  $Q$  is at least 2. Since such an event occurs with a positive probability, we obtain  $P(Q \geq 2) > 0$ . Moreover,  $P(R \geq (n-1)/2) > 0$  and by (2.40) we have  $P(R \geq (n-1)/2, Q \geq 2) = 0$ . Therefore,  $R$  and  $Q$  are dependent as well.  $\square$

*Remark 2.7.2.* In the proof of Theorem 2.7.1 we use Lemma 2.3.3 for  $k = 1$  and we also need  $n$  to be at least 9 to have  $(n-9)/2$  balls in the case of odd  $n$ . Thus, the lower bound for  $n$  is  $\max\{9, \kappa' + 3\}$ . Note that this lower bound is  $\kappa' + 3$  for  $d \geq 3$ , and it is 5 for  $d = 1$  and 8 for  $d = 2$ .

We next show that  $Q_j$ 's are pairwise dependent for large  $n$ .

**Theorem 2.7.3.** *For every  $n \geq 2\kappa' + 14$ ,  $Q_a$  and  $Q_b$  are dependent for all  $0 \leq a \neq b \leq \kappa'$ .*

*Proof.* We first prove the statement for  $1 \leq a \neq b \leq \kappa'$ . For each  $1 \leq j \leq \kappa'(d)$ , let  $t_j$  and  $s_j$  be integers such that  $n = (j+1)t_j + s_j$  and  $0 \leq s_j \leq j$  (such integers exist by the division algorithm applied to  $n$  and  $j+1$ ). We show that  $P(Q_j \geq t_j) > 0$ .

Consider the balls  $B_\epsilon(a_0), B_\epsilon(a_1), \dots, B_\epsilon(a_j)$  defined in the proof of Lemma 2.3.3. Recall that whenever there exists exactly one point in each of these balls and the remaining sample points are far enough, the indegree of the point in  $B_\epsilon(a_0)$  is  $j$ . Now consider  $t_j$  copies of this configuration and a small ball far from each other. Suppose each ball of the copies contains exactly one point and the remaining points are in the small ball. In this case, we have at least  $t_j$  points with indegree  $j$ . See Figure 2.6 for an illustration. Since having such a configuration is an event with positive probability, we obtain that  $P(Q_j \geq t_j) > 0$ .

Recall that since the sum of the indegrees in a NND is equal to the number of arcs, we have  $n = \sum_{j=0}^{\kappa'} jQ_j$ . Therefore, whenever both of the events  $\{Q_a \geq t_a\}$  and  $\{Q_b \geq t_b\}$

occur, we have

$$n = \sum_{j=0}^{\kappa'} jQ_j \geq aQ_a + bQ_b \geq at_a + bt_b > a \left( \frac{n}{a+1} - 1 \right) + b \left( \frac{n}{b+1} - 1 \right)$$

since  $t_a > \frac{n}{a+1} - 1$  and  $t_b > \frac{n}{b+1} - 1$ . Thus, we get

$$a + b > n \left( \frac{a}{a+1} + \frac{b}{b+1} - 1 \right)$$

which is equivalent to

$$n < (a+b) \frac{ab+a+b+1}{ab-1} = (a+b) \left( 1 + \frac{a+b+2}{ab-1} \right). \quad (2.41)$$

We may assume  $a > b$ . Then  $a \geq 2$  and  $ab-1 \geq b(a-1)$ , thus we have

$$\begin{aligned} (a+b) \left( 1 + \frac{a+b+2}{ab-1} \right) &\leq (a+b) \left( 1 + \frac{a+b+2}{b(a-1)} \right) \\ &= a+b + \frac{a^2+b^2+2ab+2a+2b}{b(a-1)} \\ &= a+b + \frac{a(a-1)+b^2+2(a-1)b+3(a-1)+4b+3}{b(a-1)} \\ &= a + \left( b + \frac{a}{b} \right) + \frac{b}{a-1} + 2 + \frac{3}{b} + \frac{4}{a-1} + \frac{3}{b(a-1)} \\ &\leq a + (a+1) + 1 + 2 + 3 + 4 + 3 = 2a + 14 \leq 2\kappa' + 14 \end{aligned}$$

since  $b + \frac{a}{b} \leq a+1$  and  $\frac{b}{a-1} \leq 1$ . Combining this result with the one in (2.41) gives  $n < 2\kappa' + 14$ , which is a contradiction. Thus, we have  $P(\{Q_a \geq t_a\} \cap \{Q_b \geq t_b\}) = 0$  and see that  $Q_a$  and  $Q_b$  are dependent whenever  $1 \leq a \neq b \leq \kappa'(d)$ .

Now recall that if  $n$  is even  $\{R = n/2\} \subseteq \{Q_0 = 0\}$  by (2.35) and  $P(R = n/2) > 0$ , and if  $n$  is odd  $\{R = (n-1)/2\} \subseteq \{Q_0 = 1\}$  by (2.36) and  $P(R = (n-1)/2) > 0$ . Thus, we conclude that  $P(Q_0 \leq 1) > 0$ . In addition, as

$$\sum_{i=0}^{\kappa'} Q_i = n = \sum_{i=0}^{\kappa'} iQ_i,$$

we get

$$Q_0 = \sum_{i=2}^{\kappa'} (i-1)Q_i = Q_2 + 2Q_3 + 3Q_4 + \cdots + (\kappa' - 1)Q_{\kappa'}.$$

Therefore, if  $Q_0 \leq 1$ , then  $Q_j = 0$  for all  $3 \leq j \leq \kappa'$ ,  $Q_2 = Q_0 \leq 1$ , and  $Q_1 = n - 2Q_0 \geq$

$n - 2$ , i.e.,

$$\{Q_0 \leq 1\} \subseteq \{Q_1 \geq n - 2\} \cap \{Q_2 \leq 1\} \cap \left( \bigcap_{j=3}^{\kappa'} \{Q_j = 0\} \right). \quad (2.42)$$

Whenever  $3 \leq j \leq \kappa'$ , by Lemma 2.3.3 we have  $P(Q_j \geq 1) > 0$ , and also, by (2.42) we obtain  $P(Q_0 \leq 1, Q_j \geq 1) = 0$ . Consequently,  $Q_0$  and  $Q_j$  are dependent for every  $3 \leq j \leq \kappa'$ .

Since  $n \geq 2\kappa' + 14 \geq 18$ , we get  $t_2 \geq 6$ , and therefore  $P(Q_2 \geq 6) > 0$ . Then (2.42) implies  $\{Q_0 \leq 1\} \cap \{Q_2 \geq 6\} = \emptyset$ , and so,  $Q_0$  and  $Q_1$  are dependent. As  $\sum_{i=0}^{\kappa'} Q_i = n$  and  $P(Q_2 \geq 6) > 0$ , we have  $P(Q_1 \leq n - 6) > 0$ . Then, by (2.42) we see that  $\{Q_0 \leq 1\} \cap \{Q_1 \leq n - 6\} = \emptyset$  and obtain the dependence of  $Q_0$  and  $Q_1$  as well.  $\square$

*Remark 2.7.4.* Recall that the sample size is not fixed in  $\mathcal{P}_n$ , and hence we can not apply the arguments used in this section for quantities based on  $\mathcal{P}_n$ .

## Chapter 3

 **$K$  NEAREST NEIGHBOR CONTINGENCY TABLES****3.1 Introduction**

Spatial clustering can occur for data points from one or multiple classes. In the one class case, the questions of interest are the existence (and level) and the number of clusters. However, in the case of multiple classes, clusters of mixed or same class might imply different types of interaction between the classes. For example, the spatial segregation occurs when members of a class tend to be found near members of the same class while spatial association occurs when members of a class tend to be found near members of the other class or classes. We study these patterns using quantities based on  $k$ NN digraph of data points for several classes.

Let  $m \geq 2$  and  $\mathcal{X}_1, \dots, \mathcal{X}_m$  be pairwise disjoint classes of points in  $\mathbb{R}^d$ . We call  $\mathcal{X}_j$  as class  $j$  for each  $1 \leq j \leq m$ . Let  $\mathcal{X}$  be the union of all points in the classes and consider the  $k$ NN digraph of  $\mathcal{X}$ . For each  $i$  and  $j$  (not necessarily distinct), the quantity we are interested in is the number of arcs in  $k$ NN digraph of  $\mathcal{X}$  whose tail is of class  $i$  and head is of class  $j$ , denoted  $N_{ij}$ . We can classify these counts according to base and  $k$ NN classes and present a contingency table called  $k$ NNCT. In this  $m \times m$  table, the entry at the intersection of the  $i$ -th row and  $j$ -th column is  $N_{ij}$  for all  $i, j \in \{1, \dots, m\}$ . Notice that the sum of the  $i$ -th row of a  $k$ NNCT is  $k$  times the number of class  $i$  points. See Table 3.1 for an example. Throughout this chapter, we use  $\alpha, \beta, \theta, \gamma$  for not necessarily distinct class labels and  $a, b, c, d$  for pairwise distinct class labels.

Probabilistic aspects of the entries in a  $k$ NNCT differ according to the underlying random structure. We basically consider four processes, namely, marked HPP, marked UBP, type 1 RL and type 2 RL.

Let  $\mathbf{N}_{\text{all}} = (N_{11}, N_{12}, \dots, N_{1m}, N_{21}, \dots, N_{mm})$ ,  $\mathbf{N}_{\text{diag}} = (N_{11}, N_{22}, \dots, N_{mm})$  and  $\mathbf{N}_{\text{off}} = (N_{12}, N_{13}, \dots, N_{(m-1)m})$  be the vector of all cell counts, all diagonal entries and all off-diagonal cell counts, respectively. Under each random process we consider, for class labels  $\alpha, \beta, \theta, \gamma$ , there exists a constant  $\rho(\alpha\beta, \theta\gamma)$  (depending on the process) such that

$$\frac{\mathbf{Cov}(N_{\alpha\beta}, N_{\theta\gamma})}{n} \rightarrow \rho(\alpha\beta, \theta\gamma) \quad (3.1)$$

as  $n \rightarrow \infty$ . (Notice that letting  $\sigma_{\alpha\beta}^2 = \rho(\alpha\beta, \alpha\beta)$  implies that  $\mathbf{Var}(N_{\alpha\beta})/n \rightarrow \sigma_{\alpha\beta}^2$ .) Therefore, there exist matrices  $\Sigma_{\text{all}}^{[k]}$ ,  $\Sigma_{\text{diag}}^{[k]}$  and  $\Sigma_{\text{off}}^{[k]}$  such that  $\mathbf{Cov}(\mathbf{N}_{\text{all}})/n \rightarrow \Sigma_{\text{all}}^{[k]}$ ,  $\mathbf{Cov}(\mathbf{N}_{\text{diag}})/n \rightarrow \Sigma_{\text{diag}}^{[k]}$  and  $\mathbf{Cov}(\mathbf{N}_{\text{off}})/n \rightarrow \Sigma_{\text{off}}^{[k]}$  as  $n \rightarrow \infty$ . We derive the entries of

Table 3.1:  $k$ NNCT for  $m$  classes.

|         |          | Classes  |          |          |          |          |                                      |
|---------|----------|----------|----------|----------|----------|----------|--------------------------------------|
|         |          | 1        | 2        | 3        | $\cdots$ | m        | Sum                                  |
| Classes | 1        | $N_{11}$ | $N_{12}$ | $N_{13}$ | $\cdots$ | $N_{1m}$ | $k \cdot \text{card}(\mathcal{X}_1)$ |
|         | 2        | $N_{21}$ | $N_{22}$ | $N_{23}$ | $\cdots$ | $N_{2m}$ | $k \cdot \text{card}(\mathcal{X}_2)$ |
|         | 3        | $N_{31}$ | $N_{32}$ | $N_{33}$ | $\cdots$ | $N_{3m}$ | $k \cdot \text{card}(\mathcal{X}_3)$ |
|         | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$                             |
|         | m        | $N_{m1}$ | $N_{m2}$ | $N_{m3}$ | $\cdots$ | $N_{mm}$ | $k \cdot \text{card}(\mathcal{X}_m)$ |
|         |          |          |          |          |          |          | $k \cdot \text{card}(\mathcal{X})$   |

these three matrices by computing the values of  $\rho(\alpha\beta, \theta\gamma)$  for every class labels  $\alpha, \beta, \theta, \gamma$  and underlying random structure. We also analyze asymptotic behavior of the vectors  $\mathbf{N}_{\text{all}}, \mathbf{N}_{\text{diag}}, \mathbf{N}_{\text{off}}$  and quadratic forms based on them.

### 3.2 Marked Homogeneous Poisson Process

Let  $p_1, \dots, p_m$  be positive real numbers with  $p_1 + \dots + p_m = 1$ . In this section, we assume that data points are obtained from marked HPP  $(\mathcal{P}_n, \mathcal{M})$  where  $\mathcal{M}$  is the probability space with sample space  $\{1, 2, \dots, m\}$  and  $P(\{i\}) = p_i$  for every  $i = 1, \dots, m$ . In other words,  $\mathcal{X}$  is obtained from  $\mathcal{P}_n$  and every  $x \in \mathcal{X}$  is marked with  $i$  with probability  $p_i$  independently of the other marks and points, and hence for each  $i$ ,  $\mathcal{X}_i$  is the set of points marked with  $i$ . Notice that this process is the same as letting  $\mathcal{X}_i$  to be a realization of  $\mathcal{P}_{np_i}$  for every  $i$ . However, we use the former process in the calculations and proofs.

#### 3.2.1 Means, Variances and Covariances of the Cell Counts under $(\mathcal{P}_n, \mathcal{M})$

Under  $(\mathcal{P}_n, \mathcal{M})$ , for every  $\alpha, \beta \in \{1, 2, \dots, m\}$ , we have

$$\mathbf{E}(N_{\alpha\beta}) = nkp_{\alpha}p_{\beta} + \epsilon_n, \quad (3.2)$$

where  $\epsilon_n$  is  $\mathcal{O}(n^k e^{-n})$ . Since  $\epsilon_n$  exponentially converges to 0 as  $n$  goes to infinity, we may ignore  $\epsilon_n$  in (3.2) and assume

$$\mathbf{E}(N_{\alpha\beta}) = nkp_{\alpha}p_{\beta}.$$

With respect to the concurrences between  $\alpha, \beta, \theta$  and  $\gamma$ ,  $\rho(\alpha\beta, \theta\gamma)$  is in one of the

eleven forms. In particular, for pairwise distinct class labels  $a, b, c, d$ , we have

$$\sigma_{aa}^2 = p_a^2(k + 2r + p_a(3k^2 - k - 4r + 2q) - 2p_a^2(k^2 - r + q)), \quad (3.3a)$$

$$\sigma_{ab}^2 = p_a p_b(k + p_b k(k - 1) + 2p_a q - 2p_a p_b(k^2 - r + q)), \quad (3.3b)$$

$$\rho(aa, bb) = -2p_a^2 p_b^2(k^2 - r + q), \quad (3.3c)$$

$$\rho(aa, ab) = p_a^2 p_b(2k^2 - k - 2r - 2p_a(k^2 - r + q)), \quad (3.3d)$$

$$\rho(aa, ba) = p_a^2 p_b(k^2 - 2r + 2q - 2p_a(k^2 - r + q)), \quad (3.3e)$$

$$\rho(ab, ba) = p_a p_b(2r + (p_b + p_a)(k^2 - 2r) - 2p_a p_b(k^2 - r + q)), \quad (3.3f)$$

$$\rho(aa, bc) = -2p_a^2 p_b p_c(k^2 - r + q), \quad (3.3g)$$

$$\rho(ab, ac) = p_a p_b p_c(k(k - 1) - 2p_a(k^2 - r + q)), \quad (3.3h)$$

$$\rho(ab, bc) = p_a p_b p_c(k^2 - 2r - 2p_b(k^2 - r + q)), \quad (3.3i)$$

$$\rho(ab, cb) = 2p_a p_b p_c(q - p_b(k^2 - r + q)), \quad (3.3j)$$

$$\rho(ab, cd) = -2p_a p_b p_c p_d(k^2 - r + q), \quad (3.3k)$$

where  $q = q(d, k)$  and  $r = r(d, k)$  are the constants introduced in Corollary 2.4.1 and Corollary 2.5.1, respectively. Recall that  $r$  and  $q$  only depend on the dimension  $d$  and  $k$ . Derivation of the results in (3.2) and (3.3a)-(3.3k) are given in Appendices C.1.

At first glance, it is not obvious that  $\sigma_{\alpha\beta}^2$  is positive. By (3.3a) we have

$$\begin{aligned} \sigma_{aa}^2 &= p_a^2(k + 2r + p_a(3k^2 - k - 4r + 2q) - 2p_a^2(k^2 - r + q)) \\ &= p_a^2((2r(1 - p_a) + k)(1 - p_a) + 2(q + k^2)p_a(1 - p_a) + k^2 p_a) \\ &> 0, \end{aligned}$$

as each term in the second row is positive. Recall that (3.3b) gives

$$\begin{aligned} \sigma_{ab}^2 &= p_a p_b(k + p_b k(k - 1) + 2p_a q - 2p_a p_b(k^2 - r + q)) \\ &= p_a p_b(2r p_a p_b + k(1 - p_a)(1 - p_b) + (2q - k^2 + k)p_a(1 - p_b) + k^2(p_a + p_b - 3p_a p_b)) \end{aligned}$$

and we also have  $2q - k^2 + k \geq 0$ , since  $q \geq (k^2 - k)/2$  by (2.25). Furthermore, we have

$$p_a + p_b - 3p_a p_b = (1 - p_a - p_b)(p_a + p_b) + (p_a - p_b)^2 + p_a p_b$$

which is clearly positive, and henceforth, we see that  $\sigma_{ab}^2$  is positive as well.

By the equations in (3.3a)-(3.3k), one can easily analyze correlation of the cell counts. For instance, as  $0 < r < k$ , we have  $k^2 - r + q > q > 0$  and hence, the result in (3.3g) implies that  $N_{aa}$  and  $N_{bc}$  are negatively correlated. On the other hand, correlation of  $N_{ab}$  and  $N_{cb}$  depend on the value of  $p_b$ . Equation (3.3j) yields that they are positively correlated if  $p_b < q/(k^2 - r + q)$  whereas they are negatively correlated if  $p_b > q/(k^2 - r + q)$ .

### 3.2.2 Asymptotic Distribution of the Cell Counts under $(\mathcal{P}_n, \mathcal{M})$

As stated in Remark 2.2.7,  $H$  in Theorem 2.1.1 can be replaced with any linear combination of  $N_{\alpha\beta}$ 's. Therefore, we obtain the joint asymptotic multivariate normality of the cell counts by applying Cramér-Wold device to  $\mathbf{N}_{\text{all}}/\sqrt{n}$ .

**Theorem 3.2.1.** *Under  $(\mathcal{P}_n, \mathcal{M})$ , we have*

$$\frac{\mathbf{N}_{\text{all}} - \mu_{\text{all}}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma_{\text{all}}^{[k]})$$

as  $n \rightarrow \infty$ , where  $\mu_{\text{all}} = nk(p_1^2, p_1p_2, \dots, p_1p_m, p_2p_1, \dots, p_m^2)$ .

Recall that  $\Sigma_{\text{all}}^{[k]}$  can be computed by the results in (3.3a)-(3.3k).

As an immediate consequence of Theorem 3.2.1 we obtain asymptotic normality of each cell count in a  $k$ NNCT.

**Corollary 3.2.2.** *Under  $(\mathcal{P}_n, \mathcal{M})$ , for every class labels  $\alpha$  and  $\beta$  we have*

$$\frac{N_{\alpha\beta} - nk p_{\alpha} p_{\beta}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{\alpha\beta}^2)$$

as  $n \rightarrow \infty$ , where  $\sigma_{\alpha\beta}^2$  is given in (3.3a) and (3.3b).

Note that by using asymptotic multivariate normality of all cell counts, one can obtain an asymptotically chi-squared distribution for quadratic forms based on the cell counts. The degree of freedom of the limiting chi-squared distribution is the rank of the limiting covariance matrix.

**Lemma 3.2.3.** *Under  $(\mathcal{P}_n, \mathcal{M})$ , the matrix  $\Sigma_{\text{all}}^{[k]}$  is invertible and hence, its rank is  $m^2$ .*

*Proof.* Note that it suffices to prove that  $\Sigma_{\text{all}}^{[k]}\mathbf{w} = \mathbf{0}$  holds only for  $\mathbf{w} = \mathbf{0}$ , where  $\mathbf{0}$  is the zero vector in  $\mathbb{R}^{m^2}$ . Let  $\mathbf{w}^t = (w_{11}, w_{12}, \dots, w_{mm})$  and  $\Sigma_{\text{all}}^{[k]}\mathbf{w} = \mathbf{0}$ . Then, we have

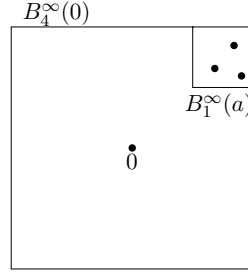
$$\mathbf{w}^t \Sigma_{\text{all}}^{[k]} \mathbf{w} = 0. \quad (3.4)$$

Now let

$$H = \sum_{\alpha=1}^m \sum_{\beta=1}^m w_{\alpha\beta} N_{\alpha\beta}. \quad (3.5)$$

By the results in (3.4) and (3.5), we obtain

$$\lim_{n \rightarrow \infty} \frac{\text{Var}(H)}{n} = 0. \quad (3.6)$$

Figure 3.1: An illustration of  $E_3$  for  $d = 2$  and  $k = 2$ .

Now suppose that  $w_{\alpha\beta} = w$  for every  $\alpha, \beta \in \{1, 2, \dots, m\}$ . Then, we get

$$H = w \sum_{\alpha=1}^m \sum_{\beta=1}^m N_{\alpha\beta} = wk \cdot \text{card}(\mathcal{X}).$$

Since  $\text{card}(\mathcal{X})$  has a Poisson distribution with parameter  $n$ , we have  $\mathbf{Var}(\text{card}(\mathcal{X})) = n$  and hence,

$$\frac{\mathbf{Var}(H)}{n} = w^2 k^2. \quad (3.7)$$

So, the results in (3.6) and (3.7) imply that  $w = 0$  and therefore, we see that  $\mathbf{w}$  is the zero vector.

We next consider the case where  $w_{\alpha\beta} \neq w_{\theta\gamma}$  for some  $\alpha, \beta, \theta$  and  $\gamma$ . Note that  $H$  is a linear combination of the cell counts and it satisfies the conditions of Proposition 2.2.1. Thus, if  $\Delta_H(\infty)$  is non-degenerate, then we get  $\sigma_H^2 > 0$  which contradicts with (3.6) and hence, we obtain the desired result. Consequently, the proof is complete if we show that  $\Delta_H(\infty)$  is non-degenerate.

Consider the events  $E_1$  and  $E_2$  defined in the proof of Corollary 2.3.4. Recall that whenever  $E_1 \cap E_2$  occurs, the insertion of a point at the origin can only affect the  $k$ NN digraph of the points in  $B_4^\infty(0)$ . Let  $a = (3, 3, \dots, 3) \in \mathbb{R}^d$  and  $E_3$  be the event that  $B_1^\infty(a)$  has exactly  $k + 1$  points from  $\mathcal{P}$  and there are no points in  $B_4^\infty(0) \setminus B_1^\infty(a)$ . See Figure 3.1 for an illustration of the event  $E_3$ . Notice that  $E_1 \cap E_2 \cap E_3$  has a positive probability. Consider any realization of  $E_1 \cap E_2 \cap E_3$ . When the point at the origin is inserted, all the  $k$ NN's of 0 are in  $B_1^\infty(a)$  and 0 is a  $k$ NN of none of the points in the sample. Therefore, insertion of 0 only adds new  $k$  arcs of the form  $(0, v)$  with  $v \in B_1^\infty(a)$  to  $k$ NN digraph.

Let  $a$  and  $b$  different class labels. If all the points in  $B_1^\infty(a)$  are marked with  $a$  and the mark of 0 is  $a$ , then  $N_{aa}$  increases by  $k$  and no other cell count changes. Thus, with

a positive probability we have

$$\Delta_H(\infty) = kw_{aa}. \quad (3.8)$$

On the other hand, if all the points in  $B_1^\infty(a)$  are marked with  $a$  and the mark of 0 is  $b$ , then  $N_{ba}$  increases by  $k$  and there happens to be no change in other cell counts. So, we get

$$\Delta_H(\infty) = kw_{ba} \quad (3.9)$$

with a positive probability. Switching the roles of  $a$  and  $b$  gives

$$\Delta_H(\infty) = kw_{bb} \text{ and } \Delta_H(\infty) = kw_{ab} \quad (3.10)$$

with a positive probability. By the assumption that  $\Delta_H(\infty)$  is a.s. constant, the results in (3.8)-(3.10) imply

$$w_{aa} = w_{ba} = w_{bb} = w_{ab}. \quad (3.11)$$

By (3.11), we see that

$$w_{\alpha\beta} = w_{\alpha\alpha} = w_{\theta\alpha} = w_{\theta\theta} = w_{\theta\gamma}$$

which contradicts with the assumption  $w_{\alpha\beta} \neq w_{\theta\gamma}$ . Therefore,  $\Delta_H(\infty)$  is non-degenerate and hence, desired result follows.  $\square$

Notice that the arguments in the proof of Lemma 3.2.3 allows to see that both  $\Sigma_{diag}^{[k]}$  and  $\Sigma_{off}^{[k]}$  are invertible as well, and thus, their ranks are  $m$  and  $(m-1)m$ , respectively. Let  $\mu_{\mathbf{diag}} = nk(p_1^2, p_2^2, \dots, p_m^2)$  and  $\mu_{\mathbf{off}} = nk(p_1p_2, p_1p_3, \dots, p_{m-1}p_m)$ . Then, combining the results in Corollary 1.1.4, Theorem 3.2.1 and Lemma 3.2.3 yield the following results.

**Proposition 3.2.4.** *Under  $(\mathcal{P}_n, \mathcal{M})$ , we have*

$$\begin{aligned} n^{-1}(\mathbf{N}_{\mathbf{all}} - \mu_{\mathbf{all}}) \left( \Sigma_{\mathbf{all}}^{[k]} \right)^{-1} (\mathbf{N}_{\mathbf{all}} - \mu_{\mathbf{all}})^t &\xrightarrow{\mathcal{L}} \chi_{m^2}^2, \\ n^{-1}(\mathbf{N}_{\mathbf{diag}} - \mu_{\mathbf{diag}}) \left( \Sigma_{\mathbf{diag}}^{[k]} \right)^{-1} (\mathbf{N}_{\mathbf{diag}} - \mu_{\mathbf{diag}})^t &\xrightarrow{\mathcal{L}} \chi_m^2, \\ n^{-1}(\mathbf{N}_{\mathbf{off}} - \mu_{\mathbf{off}}) \left( \Sigma_{\mathbf{off}}^{[k]} \right)^{-1} (\mathbf{N}_{\mathbf{off}} - \mu_{\mathbf{off}})^t &\xrightarrow{\mathcal{L}} \chi_{(m-1)m}^2, \end{aligned}$$

as  $n \rightarrow \infty$ .

*Remark 3.2.5.* One can also generalize the results in Proposition 3.2.4. Let  $T = \{(s_1, t_1), \dots, (s_l, t_l)\}$  be a nonempty subset of  $\{1, \dots, m\} \times \{1, \dots, m\}$ ,  $\mathbf{N}_T = (N_{s_1t_1}, \dots, N_{s_lt_l})$  and  $\mu_T = nk(p_{s_1p_{t_1}}, \dots, p_{s_lp_{t_l}})$ . Then,  $\mathbf{Cov}(\mathbf{N}_T)/n \rightarrow \Sigma_T^{[k]}$  as  $n \rightarrow \infty$  for some  $l \times l$

invertible constant matrix  $\Sigma_T^{[k]}$ . Moreover, we have

$$n^{-1}(\mathbf{N}_T - \mu_T) \left( \Sigma_T^{[k]} \right)^{-1} (\mathbf{N}_T - \mu_T)^t \xrightarrow{\mathcal{L}} \chi_l^2,$$

as  $n \rightarrow \infty$ .

### 3.3 Marked Uniform Binomial Process

In this section, we suppose that data points are obtained from marked UBP  $(\mathcal{U}_n, \mathcal{M})$  where  $\mathcal{M}$  is the probability space with sample space  $\{1, 2, \dots, m\}$  and  $P(\{i\}) = p_i$  for every  $i = 1, \dots, m$ . In this setting,  $\mathcal{X}$  is obtained from  $\mathcal{U}_n$  and every  $x \in \mathcal{X}$  is marked with respect to  $\mathcal{M}$  as in the case of marked HPP. Thus, for every  $i$ ,  $\mathcal{X}_i$  is the class of points marked with  $i$ . Notice that  $(\mathcal{U}_n, \mathcal{M})$  is  $(\mathcal{P}_n, \mathcal{M})$  conditioned on the sample size to be  $n$ .

It is easy to see that under  $(\mathcal{U}_n, \mathcal{M})$  we have

$$\mathbf{E}(N_{\alpha\beta}) = nkp_{\alpha}p_{\beta}.$$

for every  $\alpha, \beta$  and  $n \geq k + 1$ .

Derivation of the covariances of the cell counts is quite similar with the marked HPP. Setting  $s = n$  in equations (C.5)-(C.14) of Appendix C.1 and then following the same procedure give the desired results. The only difference in the calculation of the covariances appears in the term  $\mathbf{E}(N^2)$  in the equation (C.15) of Appendix C.1 where  $N = \text{card}(\mathcal{X})$ . Since under UBP the sample size is always  $n$ , the term  $\mathbf{E}(N^2)$  gives  $n^2$  instead of  $n^2 + n$ . Therefore, we have that  $\mathbf{Cov}(N_{\alpha\beta}, N_{\theta\gamma})$  under UBP would be  $nk^2p_{\alpha}p_{\beta}p_{\theta}p_{\gamma}$  less than in the case of marked HPP. So, under marked UBP we have

$$\sigma_{aa}^2 = p_a^2(k + 2r + p_a(3k^2 - k - 4r + 2q) - p_a^2(3k^2 - 2r + 2q)), \quad (3.12a)$$

$$\sigma_{ab}^2 = p_ap_b(k + p_bk(k - 1) + 2p_aq - p_ap_b(3k^2 - 2r + 2q)), \quad (3.12b)$$

$$\rho(aa, bb) = -p_a^2p_b^2(3k^2 - 2r + 2q), \quad (3.12c)$$

$$\rho(aa, ab) = p_a^2p_b(2k^2 - k - 2r - p_a(3k^2 - 2r + 2q)), \quad (3.12d)$$

$$\rho(aa, ba) = p_a^2p_b(k^2 - 2r + 2q - p_a(3k^2 - 2r + 2q)), \quad (3.12e)$$

$$\rho(ab, ba) = p_ap_b(2r + (p_b + p_a)(k^2 - 2r) - p_ap_b(3k^2 - 2r + 2q)), \quad (3.12f)$$

$$\rho(aa, bc) = -p_a^2p_bp_c(3k^2 - 2r + 2q), \quad (3.12g)$$

$$\rho(ab, ac) = p_ap_bp_c(k(k - 1) - p_a(3k^2 - 2r + 2q)), \quad (3.12h)$$

$$\rho(ab, bc) = p_ap_bp_c(k^2 - 2r - p_b(3k^2 - 2r + 2q)), \quad (3.12i)$$

$$\rho(ab, cb) = p_ap_bp_c(q - p_b(3k^2 - 2r + 2q)), \quad (3.12j)$$

$$\rho(ab, cd) = -p_ap_bp_cp_d(3k^2 - 2r + 2q), \quad (3.12k)$$

for pairwise distinct class labels  $a, b, c, d$ .

Note that, in a similar way of the case marked HPP, we have the joint asymptotic multivariate normality of all cell counts by Remark 2.2.7.

**Theorem 3.3.1.** *Under  $(\mathcal{U}_n, \mathcal{M})$ , we have*

$$\frac{\mathbf{N}_{\text{all}} - \mu_{\text{all}}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma_{\text{all}}^{[k]}),$$

as  $n \rightarrow \infty$ .

Recall that  $\mu_{\text{all}} = nk(p_1^2, p_1p_2, \dots, p_1p_m, p_2p_1, \dots, p_m^2)$  and entries of  $\Sigma_{\text{all}}^{[k]}$  can be computed by the results in (3.12a)-(3.12k)

Following the same idea in the proof of Lemma 3.2.3, one can easily see that the kernel of  $\Sigma_{\text{all}}^{[k]}$  is the span of  $(1, 1, \dots, 1) \in \mathbb{R}^{m^2}$  and hence, rank of  $\Sigma_{\text{all}}^{[k]}$  is  $m^2 - 1$ . On the other hand, same ideas yield that both  $\Sigma_{\text{diag}}^{[k]}$  and  $\Sigma_{\text{off}}^{[k]}$  are of full rank, and so, we obtain the following result by Corollary 1.1.4 and Theorem 3.3.1.

**Proposition 3.3.2.** *Under  $(\mathcal{U}_n, \mathcal{M})$ , we have*

$$\begin{aligned} n^{-1}(\mathbf{N}_{\text{all}} - \mu_{\text{all}}) \left( \Sigma_{\text{all}}^{[k]} \right)^+ (\mathbf{N}_{\text{all}} - \mu_{\text{all}})^t &\xrightarrow{\mathcal{L}} \chi_{m^2-1}^2, \\ n^{-1}(\mathbf{N}_{\text{diag}} - \mu_{\text{diag}}) \left( \Sigma_{\text{diag}}^{[k]} \right)^{-1} (\mathbf{N}_{\text{diag}} - \mu_{\text{diag}})^t &\xrightarrow{\mathcal{L}} \chi_m^2, \\ n^{-1}(\mathbf{N}_{\text{off}} - \mu_{\text{off}}) \left( \Sigma_{\text{off}}^{[k]} \right)^{-1} (\mathbf{N}_{\text{off}} - \mu_{\text{off}})^t &\xrightarrow{\mathcal{L}} \chi_{(m-1)m}^2, \end{aligned}$$

as  $n \rightarrow \infty$ .

### 3.4 Random Labelling

Let  $(V_n)_{n \geq 1}$  be a fixed sequence of point sets with  $\text{card}(V_n) = n$  for every  $n$ . Suppose that every point of  $V_n$  is labelled with respect to  $\mathcal{M}$  independently of the other points in the set, and similarly, points with label  $i$  are considered to be a member of the class  $i$ . Note that this process is same as conditioning on the data points in marked UBP. So, in this setting, randomness resides only on the labels (or marks) of the points and hence, we call this process as type 1 RL. If we condition on the number of class labels as well, then we obtain another random process and denote it as type 2 RL.

In type 2 RL setting, we assume that we have  $n_i := n_i(n)$  class  $i$  labels for each  $i \in \{1, \dots, m\}$  such that  $n_1 + \dots + n_m = n$ , and class identities are randomly permuted over the points so that each permutation is equally likely to occur. In other words,  $n_1$  class 1 labels,  $\dots$ ,  $n_m$  class  $m$  labels are randomly distributed over the  $n$  points in  $V_n$  in a way that each point has a unique class label. In the analysis of type 2 RL, we also suppose that there exist nonnegative real numbers  $p_1, \dots, p_m$  such that  $n_i/n \rightarrow p_i$  as  $n \rightarrow \infty$  for every  $i \in \{1, \dots, m\}$ .

Moreover, under both of the random labelling processes, we assume that

$$\frac{R^{[k]}(V_n)}{n} \rightarrow r \text{ and } \frac{Q^{[k]}(V_n)}{n} \rightarrow q, \quad (3.13)$$

as  $n \rightarrow \infty$ , for some real numbers  $r$  and  $q$ . Throughout this section,  $R^{[k]}(V_n)$  and  $Q^{[k]}(V_n)$  are denoted as  $R_n^{[k]}$  and  $Q_n^{[k]}$  for convenience in the text, respectively. We first study type 1 RL.

### 3.4.1 Type 1 RL

#### 3.4.1.1 Means, Variances and Covariances of the Cell Counts under Type 1 RL

Similar to the case of marked UBP, under type 1 RL we have

$$\mathbf{E}(N_{\alpha\beta}) = nkp_{\alpha}p_{\beta}.$$

for every  $\alpha, \beta$  and  $n \geq k + 1$ . By replacing  $s$  with  $n$  in equations (C.5)-(C.14) of Appendix C.1, one can easily obtain the covariance of the cell entries in a  $k$ NNCT under type 1 RL. The only difference is number of reflexive and shared  $k$ NN's are no longer random variables, and thus their expectations are replaced with  $R_n^{[k]}$  and  $Q_n^{[k]}$ , respectively. Therefore, under type 1 RL, for pairwise distinct class labels  $a, b, c, d$ , we get

$$\begin{aligned} \mathbf{Var}(N_{aa}) &= p_a^2(nk + 2R_n^{[k]} + p_a(3nk^2 - nk - 4R_n^{[k]} + 2Q_n^{[k]})) \\ &\quad - p_a^2(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]}), \end{aligned} \quad (3.14a)$$

$$\begin{aligned} \mathbf{Var}(N_{ab}) &= p_ap_b(nk + p_bnk(k-1) + 2p_aQ_n^{[k]} \\ &\quad - p_ap_b(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]})), \end{aligned} \quad (3.14b)$$

$$\mathbf{Cov}(N_{aa}, N_{bb}) = -p_a^2p_b^2(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]}), \quad (3.14c)$$

$$\mathbf{Cov}(N_{aa}, N_{ab}) = p_a^2p_b(2nk^2 - nk - 2R_n^{[k]} - p_a(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]})), \quad (3.14d)$$

$$\mathbf{Cov}(N_{aa}, N_{ba}) = p_a^2p_b(nk^2 - 2R_n^{[k]} + 2Q_n^{[k]} - p_a(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]})), \quad (3.14e)$$

$$\begin{aligned} \mathbf{Cov}(N_{ab}, N_{ba}) &= p_ap_b(2R_n^{[k]} + (p_a + p_b)(nk^2 - 2R_n^{[k]})) \\ &\quad - p_ap_b(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]}), \end{aligned} \quad (3.14f)$$

$$\mathbf{Cov}(N_{aa}, N_{bc}) = -p_a^2p_bp_c(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]}), \quad (3.14g)$$

$$\mathbf{Cov}(N_{ab}, N_{ac}) = p_ap_bp_c(nk(k-1) - p_a(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]})), \quad (3.14h)$$

$$\mathbf{Cov}(N_{ab}, N_{bc}) = p_ap_bp_c(nk^2 - 2R_n^{[k]} - p_b(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]})), \quad (3.14i)$$

$$\mathbf{Cov}(N_{ab}, N_{cb}) = p_ap_bp_c(2Q_n^{[k]} - p_b(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]})), \quad (3.14j)$$

$$\mathbf{Cov}(N_{ab}, N_{cd}) = -p_ap_bp_cp_d(3nk^2 - 2R_n^{[k]} + 2Q_n^{[k]}). \quad (3.14k)$$

Thus, by the assumption in (3.13) and the equations in (3.14a)-(3.14k), we have

$$\sigma_{aa}^2 = p_a^2(k + 2r + p_a(3k^2 - k - 4r + 2q) - p_a^2(3k^2 - 2r + 2q)), \quad (3.15a)$$

$$\sigma_{ab}^2 = p_a p_b(k + p_b k(k - 1) + 2p_a q - p_a p_b(3k^2 - 2r + 2q)), \quad (3.15b)$$

$$\rho(aa, bb) = -p_a^2 p_b^2(3k^2 - 2r + 2q), \quad (3.15c)$$

$$\rho(aa, ab) = p_a^2 p_b(2k^2 - k - 2r - p_a(3k^2 - 2r + 2q)), \quad (3.15d)$$

$$\rho(aa, ba) = p_a^2 p_b(k^2 - 2r + 2q - p_a(3k^2 - 2r + 2q)), \quad (3.15e)$$

$$\rho(ab, ba) = p_a p_b(2r + (p_b + p_a)(k^2 - 2r) - p_a p_b(3k^2 - 2r + 2q)), \quad (3.15f)$$

$$\rho(aa, bc) = -p_a^2 p_b p_c(3k^2 - 2r + 2q), \quad (3.15g)$$

$$\rho(ab, ac) = p_a p_b p_c(k(k - 1) - p_a(3k^2 - 2r + 2q)), \quad (3.15h)$$

$$\rho(ab, bc) = p_a p_b p_c(k^2 - 2r - p_b(3k^2 - 2r + 2q)), \quad (3.15i)$$

$$\rho(ab, cb) = p_a p_b p_c(q - p_b(3k^2 - 2r + 2q)), \quad (3.15j)$$

$$\rho(ab, cd) = -p_a p_b p_c p_d(3k^2 - 2r + 2q), \quad (3.15k)$$

under type 1 RL. Thus, the matrix  $\Sigma_{all}^{[k]}$  under type 1 RL can be computed by the equations in (3.15a)-(3.15k).

*Remark 3.4.1.* Notice that these equations are exactly the same with in marked UBP (see, equations (3.12a)-(3.12k)). However, the constants  $r$  and  $q$  in those results are equal to  $r(d, k)$  and  $q(d, k)$ , respectively, whereas under type 1 RL they are not required to be  $r(d, k)$  and  $q(d, k)$ .

### 3.4.1.2 Asymptotic Distribution of the Cell Counts under Type 1 RL

Since the point sets are fixed in random labelling setting, we can not apply subdigraph counts results under type 1 RL. So, we use another type of CLT results.

For a set of random variables  $\{Z_i, i \in V\}$  indexed by a vertices of graph  $G = (V, E)$ ,  $G$  is called a *dependency graph* if for any pair of disjoint subsets  $A_1$  and  $A_2$  of  $V$  such that there is no edge in  $E$  whose one end point is in  $A_1$  and the other is in  $A_2$ , the sets of random variables  $\{Z_i, i \in A_1\}$  and  $\{Z_i, i \in A_2\}$  are independent.

**Theorem 3.4.2** (Corollary 2 in Baldi and Rinott (1989)). *Let  $\{Z_i, i \in V\}$  be random variables having a dependency graph  $G = (V, E)$ . Set  $W = \sum_{i \in V} Z_i$  and  $\sigma^2 = \text{Var}(W)$ . Let  $D$  denote the maximal degree in  $G$  and suppose  $|Z_i| \leq B$  a.s. Define  $Q = \text{card}(V)D^2 B^3 / \sigma^3$ . Then, we have*

$$\left| P\left(\frac{W - \mathbf{E}(W)}{\sigma} \leq w\right) - \Phi(w) \right| \leq 32(1 + \sqrt{6})\sqrt{Q},$$

for any real number  $w$ , where  $\Phi$  is the cumulative distribution function of the standard normal variable.

We use Theorem 3.4.2 to prove that  $\mathbf{N}_{all}$  is asymptotically multivariate normal.

**Theorem 3.4.3.** *Under type 1 RL, we have*

$$\frac{\mathbf{N}_{all} - \mu_{all}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma_{all}^{[k]}),$$

as  $n \rightarrow \infty$ .

*Proof.* Let

$$H = \sum_{\alpha=1}^m \sum_{\beta=1}^m w_{\alpha\beta} N_{\alpha\beta}$$

be any linear combination of  $N_{\alpha\beta}$ 's and  $\mathbf{w}^t = (w_{11}, w_{12}, \dots, w_{mm}) \in \mathbb{R}^{m^2}$ . Then, note that  $\mathbf{Var}(H)/n \rightarrow \mathbf{w}^t \Sigma_{all}^{[k]} \mathbf{w}$  as  $n \rightarrow \infty$ . So, by Cramér-Wold theorem, it suffices to show that  $H$  is asymptotically normal.

If  $\mathbf{w}^t \Sigma_{all}^{[k]} \mathbf{w} = 0$ , then Chebyshev's inequality implies that  $(H - \mathbf{E}(H))/\sqrt{n}$  converges in probability to 0, and hence converges in law to 0 as well. As we may consider 0 as degenerate normal variable  $\mathcal{N}(0, 0)$ , the result follows.

Otherwise,  $\mathbf{w}^t \Sigma_{all}^{[k]} \mathbf{w} \neq 0$ . In other words,  $\mathbf{Var}(H)/n$  converges to a positive number as  $n$  goes to infinity and thus,  $\mathbf{Var}(H)$  is of order  $n$ .

Let  $v_1, \dots, v_n$  be the points in  $V_n$ . For any  $i \in \{1, \dots, n\}$ ,  $j \in \{1, \dots, k\}$  and  $\alpha, \beta \in \{1, \dots, m\}$ , let  $L_{ij}^{\alpha\beta}$  be the indicator of the event that label of  $v_i$  is  $\alpha$  and label of the  $j$ -th NN of  $v_i$  is  $\beta$ . Then, notice that

$$N_{\alpha\beta} = \sum_{j=1}^k \sum_{i=1}^n L_{ij}^{\alpha\beta}$$

and hence,

$$H = \sum_{j=1}^k \sum_{i=1}^n Z_{ij},$$

where

$$Z_{ij} = \sum_{\alpha=1}^m \sum_{\beta=1}^m w_{\alpha\beta} L_{ij}^{\alpha\beta}.$$

We next construct a dependency graph for  $Z_{ij}$ 's. Let  $V = \{(i, j) : i \in \{1, \dots, n\}, j \in \{1, \dots, k\}\}$  be the vertex set. Draw an edge between  $(i_1, j_1)$  and  $(i_2, j_2)$  whenever  $\{i_1, j_1\} \cap \{i_2, j_2\} \neq \emptyset$ . In the constructed graph, vertices correspond to the arcs in the  $k$ NN digraph and edges represent pair of arcs sharing a vertex. Clearly, this graph is a dependency graph for  $Z_{ij}$ 's. In the  $k$ NN digraph, each vertex is the tail of exactly  $k$  arcs. Also, recall that there exist a constant  $\kappa$  which only depends on  $d$  and  $k$  such

that a vertex is a  $k$ NN of at most  $\kappa$  other points and hence, the maximum degree in the dependency graph is less than  $\delta = 2k + 2\kappa$ . In other words, we have  $D < \delta$  and  $\delta$  is independent of  $n$ .

Let  $w_{max} = \max_{\alpha, \beta} |w_{\alpha\beta}|$ . Then, as  $L_{ij}^{\alpha\beta}$ 's are indicator random variables, we have

$$|Z_{ij}| \leq \sum_{\alpha=1}^m \sum_{\beta=1}^m |w_{\alpha\beta}| \leq m^2 w_{max}$$

for every  $(i, j) \in V$ . Therefore, we can set  $B = m^2 w_{max}$ . Finally, we get

$$\frac{\text{card}(V) D^2 B^3}{(\text{Var}(H))^{3/2}} < \frac{nk \delta^2 m^6 w_{max}^3}{(\text{Var}(H))^{3/2}}.$$

As  $n/(\text{Var}(H))^{3/2} \rightarrow 0$  as  $n \rightarrow \infty$ , we obtain that  $H$  is asymptotically normal by Theorem 3.4.2.  $\square$

Recall that if we set the limiting numbers  $r$  and  $q$  in (3.13) to be  $r(d, k)$  and  $q(d, k)$ , then under type 1 RL we have the same  $\Sigma_{all}^{[k]}$  obtained for marked HPP. And thus, in this case, we get the same asymptotic results for the quadratic forms.

**Proposition 3.4.4.** *Under type 1 RL, if  $r = r(d, k)$  and  $q = q(d, k)$ , then we have*

$$\begin{aligned} n^{-1}(\mathbf{N}_{all} - \mu_{all}) \left( \Sigma_{all}^{[k]} \right)^+ (\mathbf{N}_{all} - \mu_{all})^t &\xrightarrow{\mathcal{L}} \chi_{m^2-1}^2, \\ n^{-1}(\mathbf{N}_{diag} - \mu_{diag}) \left( \Sigma_{diag}^{[k]} \right)^{-1} (\mathbf{N}_{diag} - \mu_{diag})^t &\xrightarrow{\mathcal{L}} \chi_m^2, \\ n^{-1}(\mathbf{N}_{off} - \mu_{off}) \left( \Sigma_{off}^{[k]} \right)^{-1} (\mathbf{N}_{off} - \mu_{off})^t &\xrightarrow{\mathcal{L}} \chi_{(m-1)m}^2, \end{aligned}$$

as  $n \rightarrow \infty$ .

*Remark 3.4.5.* The results in Henze (1987) and Schilling (1986b) imply that proportion of expected number of reflexive pairs to  $n$  and proportion of expected number of shared  $k$ NN's to  $n$  converges to  $r(d, k)$  and  $q(d, k)$ , respectively, whenever data points are independently drawn from a distribution with an a.e. continuous density. That is,  $\mathbf{E}(R^{[k]}(\mathcal{X}_n))/n \rightarrow r(d, k)$  and  $\mathbf{E}(Q^{[k]}(\mathcal{X}_n))/n \rightarrow q(d, k)$  as  $n \rightarrow \infty$ , where  $\mathcal{X}_n$  is obtained by an  $n$  independent copies of a random variable  $X$  with an a.e. density in  $\mathbb{R}^d$ . (Note that we verified their results when  $X$  has uniform distribution over a fixed region.) Therefore, the condition  $r = r(d, k)$  and  $q = q(d, k)$  in Proposition 3.4.4 is a quite realistic assumption.

*Remark 3.4.6.* The degrees of freedom of the limiting chi-squared distributions in Proposition 3.4.4 differ with respect to the values of  $r$  and  $k$ . For example, consider the following setting: for  $n = (k+1)s_n + t_n$  where  $s_n$  and  $t_n$  are nonnegative integers with  $0 \leq t_n \leq k$ , let  $V_n$  consist of pairwise far enough  $s_n + 1$  clusters such that  $s_n$  of them contains exactly  $k+1$  points and one contains  $t_n$  points. Notice that in such a case,

members of a cluster are  $k$ NN to each other and hence, one can easily see that the number of reflexive pairs is  $s_n(k+1)k/2 + t_n(t_n+1)/2$ . As  $s_n$  is approximately  $n/(k+1)$  and  $0 \leq t_n \leq k$ , we obtain  $r = k/2$ . In other words, approximately each  $k$ NN relation is symmetric. Moreover, almost all the indegrees in the  $k$ NN digraph of  $V_n$  are  $k$ , and therefore we have  $q = k(k-1)/2$ . On the other hand, since almost all the  $k$ NN relations are symmetric,  $N_{ab}$  and  $N_{ba}$  can not differ so much. That is, there exist a constant  $c$  such that  $|N_{ab} - N_{ba}| \leq c$  and so,  $(N_{ab} - \mathbf{E}(N_{ab}))/\sqrt{n}$  and  $(N_{ba} - \mathbf{E}(N_{ba}))/\sqrt{n}$  are asymptotically equally distributed. Consequently, in such a setting, the rank of  $\Sigma_{all}^{[k]}$  is no longer  $m^2 - 1$  and in fact it is at most  $m(m+1)/2 - 1$ .

### 3.4.2 Type 2 RL

Recall that under type 2 RL we assume that  $n_1$  class 1 labels,  $\dots$ ,  $n_m$  class  $m$  labels are randomly distributed over the points of  $V_n$  and  $n_i/n \rightarrow p_i$  for each  $i \in \{1, \dots, m\}$ .

Note that under each of the previous three random structures the probability that two given points are marked or labelled with  $i$  is  $p_i^2$ . But under type 2 RL, it is  $n_i(n_i - 1)/n(n-1)$  since the number of class  $i$  points is fixed to be  $n_i$ . Particularly, class labels of the points are not independently assigned as in the other processes.

#### 3.4.2.1 Means, Variances and Covariances of the Cell Counts under Type 2 RL

For any given distinct sample points  $u, v, w, z$ , under type 2 RL, let  $p_{\alpha\beta}$  be the probability that  $u$  is of class  $\alpha$  and  $v$  is of class  $\beta$ ;  $p_{\alpha\beta\theta}$  be the probability that  $u, v$  and  $w$  have labels  $\alpha, \beta$ , and  $\theta$ , respectively; and  $p_{\alpha\beta\theta\gamma}$  be the probability that points  $u, v, w$  and  $z$  have labels  $\alpha, \beta, \theta$  and  $\gamma$ , respectively. In particular, for pairwise distinct class labels  $a, b, c, d$ , we have

$$\begin{aligned} p_{aa} &= \frac{n_a(n_a - 1)}{n(n-1)} \text{ and } p_{ab} = \frac{n_a n_b}{n(n-1)}; \\ p_{aaa} &= \frac{n_a(n_a - 1)(n_a - 2)}{n(n-1)(n-2)}, p_{aab} = \frac{n_a(n_a - 1)n_b}{n(n-1)(n-2)}, \text{ and } p_{abc} = \frac{n_a n_b n_c}{n(n-1)(n-2)}; \\ p_{aaaa} &= \frac{n_a(n_a - 1)(n_a - 2)(n_a - 3)}{n(n-1)(n-2)(n-3)}, p_{aaab} = \frac{n_a(n_a - 1)(n_a - 2)n_b}{n(n-1)(n-2)(n-3)}, \\ p_{aabb} &= \frac{n_a(n_a - 1)n_b(n_b - 1)}{n(n-1)(n-2)(n-3)}, p_{aabc} = \frac{n_a(n_a - 1)n_b n_c}{n(n-1)(n-2)(n-3)}, \\ \text{and } p_{abcd} &= \frac{n_a n_b n_c n_d}{n(n-1)(n-2)(n-3)}. \end{aligned}$$

Recall that the probabilities above are independent of  $u, v, w, z$ , and invariant under any permutation of the indices, i.e.,  $p_{ab} = p_{ba}$ ,  $p_{aab} = p_{aba} = p_{baa}$  and so on.

For every  $\alpha, \beta \in \{1, \dots, m\}$ , we have

$$\mathbf{E}(N_{\alpha\beta}) = nk p_{\alpha\beta}.$$

In particular, for every  $a, b \in \{1, \dots, m\}$  with  $a \neq b$ , we have

$$\mathbf{E}(N_{aa}) = \frac{kn_a(n_a - 1)}{n - 1} \text{ and } \mathbf{E}(N_{ab}) = \frac{kn_a n_b}{n - 1}.$$

For pairwise distinct class labels  $a, b, c, d$ , we have

$$\begin{aligned} \mathbf{Var}(N_{aa}) &= p_{aa}(nk + 2R_n^{[k]}) + p_{aaa}(3nk^2 - nk - 4R_n^{[k]} + 2Q_n^{[k]}) \\ &\quad + p_{aaaa}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - (nk p_{aa})^2, \end{aligned} \quad (3.16a)$$

$$\begin{aligned} \mathbf{Var}(N_{ab}) &= p_{ab}nk + p_{abb}nk(k - 1) + 2p_{aab}Q_n^{[k]} \\ &\quad + p_{abbb}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - (nk p_{ab})^2, \end{aligned} \quad (3.16b)$$

$$\mathbf{Cov}(N_{aa}, N_{bb}) = p_{aabb}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{aa}p_{bb}n^2k^2, \quad (3.16c)$$

$$\begin{aligned} \mathbf{Cov}(N_{aa}, N_{ab}) &= p_{aabb}(2nk^2 - nk - 2R_n^{[k]}) + p_{aaab}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{aa}p_{ab}n^2k^2, \end{aligned} \quad (3.16d)$$

$$\begin{aligned} \mathbf{Cov}(N_{aa}, N_{ba}) &= p_{aabb}(nk^2 - 2R_n^{[k]} + 2Q_n^{[k]}) + p_{aaab}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{aa}p_{ab}n^2k^2, \end{aligned} \quad (3.16e)$$

$$\begin{aligned} \mathbf{Cov}(N_{ab}, N_{ba}) &= 2p_{ab}R_n^{[k]} + (p_{abb} + p_{aab})(nk^2 - 2R_n^{[k]}) \\ &\quad + p_{abbb}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{ab}p_{ba}n^2k^2, \end{aligned} \quad (3.16f)$$

$$\mathbf{Cov}(N_{aa}, N_{bc}) = p_{aabc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{aa}p_{bc}n^2k^2, \quad (3.16g)$$

$$\begin{aligned} \mathbf{Cov}(N_{ab}, N_{ac}) &= p_{abcb}nk(k - 1) + p_{aabc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{ab}p_{ac}n^2k^2, \end{aligned} \quad (3.16h)$$

$$\begin{aligned} \mathbf{Cov}(N_{ab}, N_{bc}) &= p_{abcb}(nk^2 - 2R_n^{[k]}) + p_{abbc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{ab}p_{bc}n^2k^2, \end{aligned} \quad (3.16i)$$

$$\mathbf{Cov}(N_{ab}, N_{cb}) = 2p_{abc}Q_n^{[k]} + p_{abbc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{ab}p_{cb}n^2k^2, \quad (3.16j)$$

$$\mathbf{Cov}(N_{ab}, N_{cd}) = p_{abcd}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{ab}p_{cd}n^2k^2. \quad (3.16k)$$

Note that for  $k = 1$ , equations (3.16a)-(3.16k) coincide with the results in Dixon (2002). Moreover, for  $m = 2$ , equations (3.16a)-(3.16f) agree with the results in Cuzick and Edwards (1990).

At first glance, variances and covariances of the cell counts might look like of order  $n^2$ . But, in fact, all of them are of order  $n$ . In each of the equations in (3.16a)-(3.16k), the coefficient of  $n^2$  is of the form  $k^2(p_{\alpha\beta\theta\gamma} - p_{\alpha\beta}p_{\theta\gamma})$ , and for any choice of  $\alpha, \beta, \theta$  and

$\gamma$ , it can be easily verified that  $p_{\alpha\beta\theta\gamma} - p_{\alpha\beta}p_{\theta\gamma}$  is  $O(1/n)$ . For example,

$$\begin{aligned} p_{aabb} - p_{aa}p_{bb} &= \frac{n_a(n_a-1)n_b(n_b-1)}{n(n-1)(n-2)(n-3)} - \frac{n_a(n_a-1)}{n(n-1)} \frac{n_b(n_b-1)}{n(n-1)} \\ &= \frac{n_a(n_a-1)n_b(n_b-1)}{n(n-1)} \left( \frac{1}{(n-2)(n-3)} - \frac{1}{n(n-1)} \right) \\ &= \frac{n_a(n_a-1)n_b(n_b-1)(4n-6)}{n^2(n-1)^2(n-2)(n-3)} \end{aligned}$$

is  $O(1/n)$ , since the numerator is  $O(n^5)$  and the denominator is  $O(n^6)$ . Furthermore,  $n(p_{aabb} - p_{aa}p_{bb}) \rightarrow 4p_a^2p_b^2$  as  $n \rightarrow \infty$ . In particular, for pairwise distinct class labels  $a, b, c, d$ , we have

$$\begin{aligned} \sigma_{aa}^2 &= p_a^2(k+2r-p_a(k^2+k+4r-2q)) + p_a^2(k^2+2r-2q) \\ &= p_a^2(1-p_a)(p_a(2q-k^2+k) + (1-p_a)(k+2r)), \end{aligned} \quad (3.17a)$$

$$\begin{aligned} \sigma_{ab}^2 &= p_ap_b(k-p_bk+p_a(2q-k^2) + p_ap_b(k^2+2r-2q)) \\ &= p_ap_b(p_a(1-p_b)(2q-k^2+k) + (1-p_a)(1-p_b)k + 2p_ap_br), \end{aligned} \quad (3.17b)$$

$$\rho(aa, bb) = p_a^2p_b^2(k^2+2r-2q), \quad (3.17c)$$

$$\rho(aa, ab) = p_a^2p_b(-k-2r+p_a(k^2+2r-2q)), \quad (3.17d)$$

$$\rho(aa, ba) = -p_a^2p_b(1-p_a)(k^2+2r-2q), \quad (3.17e)$$

$$\rho(ab, ba) = p_ap_b(2(1-p_a-p_b)r + p_ap_b(k^2+2r-2q)), \quad (3.17f)$$

$$\rho(aa, bc) = p_a^2p_bp_c(k^2+2r-2q), \quad (3.17g)$$

$$\rho(ab, ac) = p_ap_bp_c(-k+p_a(k^2+2r-2q)), \quad (3.17h)$$

$$\rho(ab, bc) = p_ap_bp_c(-2r+p_b(k^2+2r-2q)), \quad (3.17i)$$

$$\rho(ab, cb) = p_ap_bp_c(q-k^2+p_b(k^2+2r-2q)), \quad (3.17j)$$

$$\rho(ab, cd) = p_ap_bp_cp_d(k^2+2r-2q), \quad (3.17k)$$

By the result in (2.24), we have  $2q-k^2+k \geq 0$  and hence, one can easily see that both  $\sigma_{ab}^2$  and  $\sigma_{ab}^2$  are positive.

*Remark 3.4.7.* The term  $k^2+2r-2q$  in equations (3.17a)-(3.17k) can attain both negative and positive values. The construction given in Remark 3.4.6 gives  $k^2+2r-2q = k^2+k-(k^2-k) = 2k > 0$ . On the other hand, consider the following setting for  $d=2$  and  $k=1$ : for  $n=6s_n+t_n$  where  $s_n$  and  $t_n$  are nonnegative integers with  $0 \leq t_n \leq 5$ , let  $V_n$  consist of pairwise far enough  $s_n+1$  clusters such that  $s_n$  of them contains exactly 6 points and one contains  $t_n$  points. Suppose that, in each of the clusters containing six points, five of the points are close to the vertices of a regular pentagon and the remaining point is close to the center of the pentagon. In each such cluster, the point near the center is the NN of other five points and hence, there is exactly one reflexive pair and ten shared

NN's. Since  $0 \leq t_n \leq 5$  and  $s_n$  is approximately  $n/6$ , we get  $r = 1/6$  and  $q = 10/6 = 5/3$ . Therefore, under this setting, we have  $k^2 + 2r - 2q = 1 + 1/3 - 10/3 = -2 < 0$ . However, if we set  $r = r(d, k)$  and  $q = q(d, k)$ , then we obtain positive values for  $k^2 + 2r - 2q$ .

### 3.4.2.2 Asymptotic Distributions of Diagonal Cell Counts under Type 2 RL

The cell counts for the cells along the diagonal have asymptotically normal distributions when  $m = 2$  (Cuzick and Edwards (1990)). But considering all the classes other than  $\alpha$  as one class allows us to apply the result for  $m = 2$ , and we see that  $N_{\alpha\alpha}$  is asymptotically normal for  $m > 2$  as well.

Schilling (1986a) and Henze (1988) consider the two-sample case and show that  $N_{11} + N_{22}$  is asymptotically normal. Henze (1988) also states that in the multi-sample case, the sum of the diagonal cell counts has an asymptotically normal distribution as well. We consider the joint distribution of the diagonal cell counts and show that it is asymptotically multivariate normal which implies all the results above.

**Theorem 3.4.8.** *Under type 2 RL, we have*

$$\frac{\mathbf{N}_{\text{diag}} - \mathbf{E}(\mathbf{N}_{\text{diag}})}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\text{diag}}^{[k]})$$

as  $n \rightarrow \infty$ .

*Proof.* By Cramér-Wold device, it suffices to prove that

$$\sum_{i=1}^m w_i N_{ii}$$

is asymptotically normal for every  $\mathbf{w} = (w_1, \dots, w_m) \in \mathbb{R}^m$ . When  $\mathbf{w}$  is the zero vector, then the claim is trivial. Therefore, we assume that  $\mathbf{w}$  is a nonzero vector.

Let  $V_n = \{v_1, \dots, v_n\}$  and

$$y_{ij} = \begin{cases} 1 & \text{if } v_j \text{ is a } k\text{NN to } v_i, \\ 0 & \text{otherwise;} \end{cases}$$

for every  $i, j \in \{1, \dots, n\}$ . Now fix a labelling of  $v_1, \dots, v_n$  and let  $L_i$  denote the class label of  $v_i$ . Set

$$b_{ij} = \sum_{a=1}^m w_a \mathbf{1}_{\{L_i=L_j=a\}},$$

for every  $i, j \in \{1, \dots, n\}$ . Notice that if  $L_i \neq L_j$ , then  $b_{ij} = 0$ , and if  $L_i = L_j = a$ , then

$b_{ij} = w_a$ . Then it is easy to see that

$$U := U(\pi) := \sum_{1 \leq i \neq j \leq n} y_{ij} b_{\pi(i)\pi(j)} \stackrel{d}{=} \sum_{i=1}^m w_i N_{ii}, \quad (3.18)$$

where  $\pi$  is uniformly distributed over all permutations of  $\{1, \dots, n\}$ . Consequently, it suffices to prove the asymptotic normality of  $U$ . Since  $b_{ij} = b_{ji}$  for every  $i$  and  $j$  (i.e., the matrix  $(b_{ij})$  is symmetric), we can rewrite  $U$  as

$$\sum_{1 \leq i \neq j \leq n} a_{ij} b_{\pi(i)\pi(j)}, \quad (3.19)$$

where  $a_{ij} = (y_{ij} + y_{ji})/2$ . (We can also symmetrize the matrix  $(y_{ij})$ .) Notice that if  $v_i$  and  $v_j$  are  $k$ NN of each other then  $a_{ij} = 1$ , if only one of them is a  $k$ NN to the other one then  $a_{ij} = 1/2$ , and if none of them is a  $k$ NN of the other one then  $a_{ij} = 0$ .

To prove that  $U$  is asymptotically normally distributed, we use Corollary 2.1 in Barbour and Eagleson (1986). Let  $(n)_i$  denote  $n(n-1)\cdots(n-i+1)$  and

$$\begin{aligned} a_1 &= \frac{1}{(n)_2} \sum_{1 \leq i \neq j \leq n} |a_{ij}|, & a_2 &= \frac{1}{(n)_3} \sum_{1 \leq i \neq j \neq k \leq n} |a_{ij} a_{ik}|, \\ a_3 &= \frac{1}{(n)_4} \sum_{1 \leq i \neq j \neq k \neq l \leq n} |a_{ij} a_{ik} a_{il}|, & a_4 &= \frac{1}{(n)_4} \sum_{1 \leq i \neq j \neq k \neq l \leq n} |a_{ij} a_{ik} a_{jl}|, \\ a_5 &= \frac{1}{(n)_3} \sum_{1 \leq i \neq j \neq k \leq n} |a_{ij}^2 a_{ik}|, & a_6 &= \frac{1}{(n)_2} \sum_{1 \leq i \neq j \leq n} |a_{ij}|^3, \\ a_7 &= \frac{1}{(n)_3} \sum_{1 \leq i \neq j \neq k \leq n} |a_{ij} a_{ik} a_{jk}|, & a_8 &= \frac{1}{(n)_2} \sum_{1 \leq i \neq j \leq n} a_{ij}^2. \end{aligned}$$

Similarly, define  $b_1, \dots, b_8$  for  $b_{ij}$ 's, and let  $s = \sqrt{\mathbf{Var}(U)}$  and

$$\epsilon = s^{-3}(n^4(a_1^3 + a_1 a_2 + a_3 + a_4)(b_1^3 + b_1 b_2 + b_3 + b_4) + n^3(a_5 + a_1 a_8)(b_5 + b_1 b_8) + n^2 a_6 b_6).$$

Corollary 2.1 in Barbour and Eagleson (1986) states that if  $\epsilon \rightarrow 0$  as  $n \rightarrow \infty$ , then  $U$  is asymptotically normal. (Notice that  $a_7$  and  $b_7$  do not appear in  $\epsilon$ , but these quantities are used in the proof of Theorem 2.1 in Barbour and Eagleson (1986) and hence, to be consistent with their notation we define these quantities as well.)

We first show that  $s/\sqrt{n}$  converges to a positive number as  $n \rightarrow \infty$ . Observe that

$$\lim_{n \rightarrow \infty} \frac{\mathbf{Var}(U)}{n} = \lim_{n \rightarrow \infty} \frac{\mathbf{Var}(\sum_{i=1}^m w_i N_{ii})}{n} = \mathbf{w} \left( \lim_{n \rightarrow \infty} \frac{\mathbf{Cov}(\mathbf{N}_{\text{diag}})}{n} \right) \mathbf{w}^t = \mathbf{w} \Sigma_{\text{diag}}^{[k]} \mathbf{w}^t. \quad (3.20)$$

If  $\mathbf{w} \Sigma_{\text{diag}}^{[k]} \mathbf{w}^t = 0$ , then  $(U - \mathbf{E}(U))/\sqrt{n}$  converges in law to the degenerate normal distribution  $\mathcal{N}(0, 0)$  and thus the claim is trivial. Otherwise, we have  $\mathbf{w} \Sigma_{\text{diag}}^{[k]} \mathbf{w}^t \neq$

0 which yields  $w \Sigma_{diag}^{[k]} w^t > 0$ , since  $\Sigma_{diag}^{[k]}$  is positive semi-definite by the symmetry. Therefore, (3.20) implies that  $s/\sqrt{n}$  converges to a positive number as  $n \rightarrow \infty$ .

Let  $w_{max} = \max_{1 \leq i \leq m} |w_i|$ . Then observe that  $0 \leq b_i \leq \max\{1, w_{max}^3\}$  for every  $i = 1, 2, \dots, 8$ .

Recall that a point is a  $k$ NN to at most  $k\kappa(d)$  other points. Therefore, there exists a constant  $K$  such that for every fixed  $i$  we have  $a_{ij} \neq 0$  for at most  $K$  different  $j$ . Moreover, as  $a_{ij} \in \{1, 1/2\}$  whenever  $a_{ij} \neq 0$ , there exist constants  $K_1, K_2$  and  $K_3$  (independent of  $n$ ) such that

$$a_1, a_6, a_8 \leq \frac{K_1}{n}, \quad a_2, a_5 \leq \frac{K_2}{n^2} \quad \text{and} \quad a_3, a_4 \leq \frac{K_3}{n^3}.$$

Then we see that  $\epsilon \leq Bn/s^3$  for some constant  $B$  and hence, we get  $\epsilon \rightarrow 0$  as  $n \rightarrow \infty$  as desired.  $\square$

By using Gerschgorin's theorem we show that, under some conditions on  $m$  and  $q$ ,  $\Sigma_{diag}^{[k]}$  is invertible.

**Theorem 3.4.9** (Gerschgorin's Theorem). *Let  $A$  be a complex  $n \times n$  matrix. For  $i \in \{1, \dots, n\}$ , let  $r_i = \sum_{j \neq i} |A_{ij}|$  be the sum of the absolute values of non-diagonal entries in the  $i$ -th row of  $A$ , and  $D(A_{ii}, r_i)$  be the closed disc centered at  $A_{ii}$  with radius  $r_i$ . Then, every eigenvalue of  $A$  lies within at least one of the discs  $D(A_{ii}, r_i)$ .*

An immediate consequence of Gerschgorin's theorem is as follows.

**Corollary 3.4.10.** *Let  $A$  and  $r_i$ 's be as defined in Theorem 3.4.9. If  $|A_{ii}| > r_i$  for each  $i \in \{1, \dots, n\}$ , then 0 is not an eigenvalue of  $A$  and hence,  $A$  is invertible.*

**Lemma 3.4.11.** *Under type 2 RL, if  $m \neq 2$  or  $q \neq (k^2 - k)/2$ , then  $\Sigma_{diag}^{[k]}$  is invertible.*

*Proof.* For convenience in the notation, we denote  $\Sigma_{diag}^{[k]}$  as  $A$  along the proof. By the results in (3.17a) and (3.17c), we have

$$A_{ii} = p_i^2(1 - p_i)(s(1 - p_i) + tp_i)$$

for each  $i \in \{1, \dots, m\}$ , and

$$A_{ij} = (s - t)p_i^2 p_j^2$$

for every  $i, j \in \{1, \dots, m\}$  with  $i \neq j$ , where  $s = k + 2r$  and  $t = 2q + k - k^2$ . Clearly we have  $s > 0$  and  $t \geq 0$  by (2.24).

Note that the condition  $q \neq (k^2 - k)/2$  is equivalent to  $t > 0$ . We first study the case

$t \leq 2s$ . The proof is based on Gerschgorin's theorem. For fixed  $i$ , we have

$$\sum_{\substack{j=1 \\ j \neq i}}^m |A_{ij}| = p_i^2 |s - t| \sum_{\substack{j=1 \\ j \neq i}}^m p_j^2 \leq p_i^2 |s - t| \left( \sum_{\substack{j=1 \\ j \neq i}}^m p_j \right)^2 = p_i^2 (1 - p_i)^2 |s - t|, \quad (3.21)$$

since  $p_i > 0$  for each  $i = 1, \dots, m$ . Now suppose  $t < s$ . Then, we have

$$\begin{aligned} p_i^2 (1 - p_i)^2 |s - t| &= p_i^2 (1 - p_i)^2 (s - t) \leq A_{ii} = p_i^2 (1 - p_i)(s(1 - p_i) + tp_i) \\ &\Leftrightarrow 0 \leq tp_i^2 (1 - p_i), \end{aligned} \quad (3.22)$$

and hence the results in (3.21) and (3.22) imply

$$\sum_{\substack{j=1 \\ j \neq i}}^m |A_{ij}| \leq A_{ii}, \quad (3.23)$$

for every  $i = 1, \dots, m$ . Notice that in (3.21) equality holds if and only if  $m = 2$ , and in (3.22) equality holds if and only if  $t = 0$ . Therefore, since  $m \neq 2$  or  $t \neq 0$ , we have a strict inequality in (3.23). Then, Gerschgorin's theorem implies that  $A$  is nonsingular.

On the other hand, if  $s \leq t \leq 2s$ , then we obtain

$$\begin{aligned} p_i^2 (1 - p_i)^2 |s - t| &= p_i^2 (1 - p_i)^2 (t - s) < A_{ii} = p_i^2 (1 - p_i)(s(1 - p_i) + tp_i) \\ &\Leftrightarrow 0 < p_i^2 (1 - p_i)((2s - t) + 2p_i(t - s)). \end{aligned} \quad (3.24)$$

As both  $2s - t$  and  $t - s$  are nonnegative and at least one of them is positive, the inequality in (3.24) is true. Thus, combining the results in (3.21) and (3.24) gives

$$\sum_{\substack{j=1 \\ j \neq i}}^m |A_{ij}| < A_{ii},$$

for every  $i = 1, \dots, m$ , and hence, again by Gerschgorin's theorem we see that  $A$  is invertible.

We now prove the claim for  $t > 2s$ . Suppose that  $A\mathbf{v} = \mathbf{0}$  for some nonzero vector  $\mathbf{v} = (v_1, \dots, v_m)$ , where  $\mathbf{0}$  is the zero vector in  $\mathbb{R}^m$ . Then for each  $i$  we have

$$p_i^2 (1 - p_i)(s(1 - p_i) + tp_i)v_i + \sum_{\substack{j=1 \\ j \neq i}}^m (s - t)p_i^2 p_j^2 v_j = 0$$

which implies

$$(s + (t - 2s)p_i)v_i = \sum_{j=1}^m (t - s)p_j^2 v_j \quad (3.25)$$

for every  $i \in \{1, \dots, m\}$ . Notice that the right hand side of (3.25) is independent of  $i$  and  $s + (t - 2s)p_i > 0$ . Then by letting  $u = t - 2s$  and  $w = \sum_{j=1}^m (t - s)p_j^2 v_j$ , equation (3.25) becomes

$$v_i = \frac{w}{s + up_i}, \quad (3.26)$$

for each  $i = 1, \dots, m$ . Then, we obtain

$$w = \sum_{j=1}^m (t - s)p_j^2 v_j = w(t - s) \sum_{j=1}^m \frac{p_j^2}{s + up_j}. \quad (3.27)$$

Recall that  $v_i \neq 0$  for some  $i$  and hence, we get  $w \neq 0$  and (3.27) implies

$$1 = (t - s) \sum_{j=1}^m \frac{p_j^2}{s + up_j}. \quad (3.28)$$

For any  $x, y > 0$  we have

$$\begin{aligned} \frac{x^2}{s + ux} + \frac{y^2}{s + uy} &< \frac{(x + y)^2}{s + u(x + y)} \\ \Leftrightarrow (s(x^2 + y^2) + uxy(x + y)) (s + u(x + y)) &< (x + y)^2 (s^2 + su(x + y) + u^2 xy) \\ \Leftrightarrow 0 &< 2s^2 xy + suxy(x + y). \end{aligned} \quad (3.29)$$

Since the last inequality in (3.29) holds, so does the first one. Then inductively we get

$$\sum_{j=1}^m \frac{p_j^2}{s + up_j} < \frac{(\sum_{j=1}^m p_j)^2}{s + u \sum_{j=1}^m p_j} = \frac{1}{s + u} = \frac{1}{t - s}. \quad (3.30)$$

Finally, observe that (3.30) contradicts with (3.28) and therefore,  $A$  is invertible.  $\square$

*Remark 3.4.12.* Notice that if both  $m = 2$  and  $q = (k^2 - k)/2$  hold, then we have

$$\Sigma_{diag}^{[k]} = sp_1^2 p_2^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

and hence,  $\Sigma_{diag}^{[k]}$  is singular and its rank is 1. Recall that if  $Q_n^{[k]} = n(k^2 - k)/2$ , then each indegree is  $k$  and hence we get  $N_{11} + N_{21} = kn_1$  and  $N_{12} + N_{22} = kn_2$ . As we also have  $N_{11} + N_{12} = kn_1$  and  $N_{21} + N_{22} = kn_2$ , we see that  $N_{22} = N_{11} - k(n_1 - n_2)$ . In other words, diagonal cell counts differ by only a constant. As an example for  $q = (k^2 - k)/2$ ,

see the setting in Remark 3.4.6.

By Corollary 1.1.4, Theorem 3.4.8 and Lemma 3.4.11, we obtain the following result.

**Proposition 3.4.13.** *Under type 2 RL, if  $m \neq 2$  or  $q \neq (k^2 - k)/2$ , then*

$$n^{-1}(\mathbf{N}_{\text{diag}} - \mathbf{E}(\mathbf{N}_{\text{diag}}))(\Sigma_{\text{diag}}^{[k]})^{-1}(\mathbf{N}_{\text{diag}} - \mathbf{E}(\mathbf{N}_{\text{diag}}))^t \xrightarrow{\mathcal{L}} \chi_m^2$$

as  $n \rightarrow \infty$ . If  $m = 2$  and  $q = (k^2 - k)/2$ , then the result holds when the inverse is replaced by the generalized inverse and the limiting distribution is  $\chi_1^2$ .

### 3.4.2.3 Off-Diagonal Cell Counts

We now consider the distribution of  $N_{ab}$ 's with  $a \neq b$ . Off-diagonal cell counts can be written in the form of in (3.19) as well. As in the previous subsection, let

$$y_{ij} = \begin{cases} 1 & \text{if } v_j \text{ is a } k\text{NN to } v_i, \\ 0 & \text{otherwise ;} \end{cases}$$

for every  $i, j \in \{1, \dots, n\}$ . For a fixed labelling of  $v_1, \dots, v_n$ , let  $L_i$  denote the class label of  $v_i$ . Then, set

$$b_{ij} = \mathbf{1}_{\{L_i=a, L_j=b\}},$$

for every  $i, j \in \{1, \dots, n\}$  and let  $\pi$  be uniformly distributed over all permutations of  $\{1, \dots, n\}$ . Then,  $N_{ab}$  is equally distributed with

$$\sum_{1 \leq i \neq j \leq n} y_{ij} b_{\pi(i)\pi(j)}. \quad (3.31)$$

A statistic of the form in (3.31) is defined as a generalized correlation coefficient (Daniels (1944); Friedman and Rafsky (1983)), random association of matrices (Barbour and Eagleson (1986)), double-indexed permutation statistics (Pham et al. (1989); Zhao et al. (1997)) and matrix correlation statistics (Barbour and Chen (2005)). Asymptotic results of this type of statistics are usually for symmetric or skew-symmetric matrices. Since the matrices  $(y_{ij})$  and  $(b_{ij})$  are neither symmetric nor skew-symmetric, we can not apply the results in Barbour and Eagleson (1986), Daniels (1944) and Pham et al. (1989).

Using Stein's method, Barbour and Chen (2005) and Zhao et al. (1997) provide Berry-Esseen type theorems for statistics of the form

$$\sum_{1 \leq i \neq j \leq n} c(i, j, \pi(i), \pi(j)),$$

where  $(c(i, j, s, t))$  is a 4-tuple real array. However, letting  $c(i, j, s, t) = y_{ij} b_{st}$  in these theorems only give constant upper bounds for the error of normal approximation, and hence, it does not yield the asymptotic normality of (3.31). Another approach to prove

the asymptotic normality of  $N_{ab}$ 's would be the method of Cuzick and Edwards (1990). But, their technique does not generalize to off-diagonal cell counts, because of the lack of symmetry. The statistics in (3.31) satisfies the conditions given in Friedman and Rafsky (1983). However, Pham et al. (1989) proves that those conditions are not sufficient for the asymptotic normality and they provide a CLT result with some extra conditions on the statistics. But our statistics do not satisfy all of the extra conditions given in Pham et al. (1989).

Yet, our extensive Monte Carlo simulation results strongly suggest that the distribution of the off-diagonal cell counts are asymptotically normal.

*Conjecture 3.4.14.* Under type 2 RL,

$$\frac{N_{ab} - \frac{kn_a n_b}{n(n-1)}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{ab}^2)$$

as  $n \rightarrow \infty$ , where  $\sigma_{ab}^2$  is given in (3.17b).

We also conjecture that  $\mathbf{N}_{\text{off}}$  and  $\mathbf{N}_{\text{all}}$  follow similar asymptotic results for  $\mathbf{N}_{\text{diag}}$ . That is, we claim that whenever  $m \neq 2$  or  $q \neq (k^2 - k)/2$ , we have

$$\Sigma_{\text{off}}^{[k]} \text{ is invertible,}$$

and hence its rank is  $m(m-1)$ ,

$$\text{the rank of } \Sigma_{\text{all}}^{[k]} \text{ is } m(m-1),$$

$$n^{-1}(\mathbf{N}_{\text{off}} - \mathbf{E}(\mathbf{N}_{\text{off}}))(\Sigma_{\text{off}}^{[k]})^{-1}(\mathbf{N}_{\text{off}} - \mathbf{E}(\mathbf{N}_{\text{off}}))^t \xrightarrow{\mathcal{L}} \chi_{m(m-1)}^2 \quad (3.32)$$

and

$$n^{-1}(\mathbf{N}_{\text{all}} - \mathbf{E}(\mathbf{N}_{\text{all}}))(\Sigma_{\text{all}}^{[k]})^+(\mathbf{N}_{\text{all}} - \mathbf{E}(\mathbf{N}_{\text{all}}))^t \xrightarrow{\mathcal{L}} \chi_{m(m-1)}^2, \quad (3.33)$$

as  $n \rightarrow \infty$ . Notice that the statements in (3.32) and (3.33) are equivalent since

$$N_{aa} = kn_a - \sum_{b=1, b \neq a}^m N_{ab},$$

for every  $a = 1, \dots, m$ .

We next provide some results on the covariance matrix of the off-diagonal cell counts.

**Lemma 3.4.15.** Suppose that  $Q_n^{[k]} \neq nk(k-1)/2$  and  $n$  is large enough. Let  $w_{ab}$  be a

real number for every  $a, b \in \{1, 2, \dots, m\}$  with  $a \neq b$ . If

$$\sum_{1 \leq a \neq b \leq m} w_{ab} N_{ab} \quad (3.34)$$

is constant, then  $w_{ab} = 0$  for every  $a$  and  $b$ .

*Proof.* Recall that  $Q_n^{[k]} = nk(k-1)/2$  if and only if all the indegrees are equal in the  $k$ NN digraph of  $V_n$ . Hence, by the assumption, there exist two points in  $V_n$  with different indegrees, say  $v_i$  and  $v_j$ . Without loss of generality, we may assume  $d_{in}(v_i) > d_{in}(v_j)$ . Let  $A$  be the set of all points  $v \in V_n$  other than  $v_i$  and  $v_j$  such that at least one of  $(v, v_i), (v, v_j), (v_i, v), (v_j, v)$  is an arc in  $k$ NN digraph of  $V_n$ ; that is, the points which are a  $k$ NN to  $v_i$  or  $v_j$  and the points in which  $v_i$  or  $v_j$  is a  $k$ NN.

Now consider a permutation of the labels such that  $v_i$  and all the points in  $A$  are of class  $a$  and  $v_j$  has label class  $b$ , where  $a \neq b$ . By the bound on the indegree of a point in a  $k$ NND, the cardinality of  $A$  is bounded above by a constant independent of  $n$ . Thus, since  $n_a$  goes to infinity as  $n$  goes to infinity, for large  $n$  such a permutation exists. If the sum in (3.34) is constant, switching the labels of two points should not change the sum. Now, switch the labels of  $v_i$  and  $v_j$ . Note that after the switch,  $N_{aa}, N_{ba}$  and  $N_{ab}$  are the only cell counts that might have changed. Clearly,  $N_{aa}$  does not affect the sum in (3.34), and  $N_{ba}$  stays fixed since both  $v_i$  and  $v_j$  are of outdegree  $k$ . However,  $N_{ab}$  increases by  $d_{in}(v_i) - d_{in}(v_j)$  which is nonzero by the choice of  $v_i$  and  $v_j$ . Therefore,  $w_{ab} = 0$  and hence, the result follows.  $\square$

**Proposition 3.4.16.** *Under type 2 RL, if  $Q_n^{[k]} \neq nk(k-1)/2$ , then  $\mathbf{Cov}(\mathbf{N}_{\text{off}})$  is invertible for large  $n$ .*

*Proof.* Let  $\mathbf{Cov}(\mathbf{N}_{\text{off}})w = 0$  for some vector  $w$  with  $w^t = (w_{12}, w_{13}, \dots, w_{m(m-1)}) \in \mathbb{R}^{m(m-1)}$ . To prove the claim, it suffices to show that  $w$  has to be the zero vector. It is easy to verify that

$$0 = w^t \mathbf{Cov}(\mathbf{N}_{\text{off}})w = \mathbf{Var} \left( \sum_{1 \leq a \neq b \leq m} w_{ab} N_{ab} \right).$$

As the variance of a discrete random variable is zero if and only if the random variable is constant, we obtain that  $\sum_{1 \leq a \neq b \leq m} w_{ab} N_{ab}$  is constant. Therefore, the result follows by Lemma 3.4.15.  $\square$

As  $N_{aa} = kn_a - \sum_{b=1, b \neq a}^m N_{ab}$  for every  $1 \leq a \leq m$ , we clearly see that  $\mathbf{Cov}(\mathbf{N}_{\text{off}})$  and  $\mathbf{Cov}(\mathbf{N}_{\text{all}})$  have the same ranks. Thus, by Proposition (3.4.16) we have the following corollary.

**Corollary 3.4.17.** *Under type 2 RL, if  $Q_n^{[k]} \neq nk(k-1)/2$ , then the rank of  $\mathbf{Cov}(\mathbf{N}_{\text{all}})$  is  $m(m-1)$  for large  $n$ .*

*Remark 3.4.18.* The condition  $Q_n^{[k]} \neq nk(k-1)/2$  is crucial for the results above. Because, for the setting given in Remark 3.4.6 if  $n$  is a multiple of  $k+1$ , then the  $k$ NN relation is always symmetric. Therefore,  $N_{ab} = N_{ba}$  for every  $a, b$  and hence, Lemma 3.4.15 fails to hold.

*Remark 3.4.19.  $S$ -NN Digraph and  $S$ -NNCT:* One can easily generalize  $k$ NN digraphs and  $k$ NNCT cell counts. Let  $S$  be a fixed finite non-empty set of positive integers. For any point set  $V$ , consider the digraph obtained by inserting the arc with tail  $u$  and head  $v$  whenever  $v$  is the  $s$ -th NN of  $u$  such that  $s \in S$ . Denote this digraph by  $D_{nn}^S(V)$ . Note that when  $S = \{1, \dots, k\}$ , we obtain  $D_{nn}^{[k]}(V)$ . For a fixed finite weakly connected digraph  $D$ , let  $H_D^S(V)$  count the copies of  $D$  in  $D_{nn}^S(V)$ . Then,  $H_D^S$  and any linear combination of such functionals satisfy the conditions of Proposition 2.2.1 and Proposition 2.2.5, and hence all the asymptotic results in the Main Theorem are valid for them as well. Let  $R^S(V)$  be the number of pairs  $\{u, v\}$  in  $V$  such that  $v$  is the  $s_1$ -th NN of  $u$  and  $u$  is the  $s_2$ -th NN of  $v$  for some  $s_1, s_2 \in S$ . Let  $Q^S(V)$  be the number of pairs of arcs in  $D_{nn}^S(V)$  sharing their heads and  $Q_j^S(V)$  be the number of vertices in  $D_{nn}^S(V)$  with indegree  $j$ . We have the same asymptotic results for  $R^S, Q^S$  and  $Q_j^S$ 's. So, there exist constants  $r(d, S)$  and  $q(d, S)$  such that  $\mathbf{E}(R^S(\mathcal{U}_n))/n \xrightarrow{c.m.c.c.} r(d, S)$  and  $\mathbf{E}(Q^S(\mathcal{U}_n))/n \xrightarrow{c.m.c.c.} q(d, S)$  as  $n \rightarrow \infty$ . Under the same setting defined for the cell counts of  $k$ NNCT for  $m$  classes, let  $N_{\alpha\beta}^S$  denote the number of arcs in  $D_{nn}^S(\mathcal{X})$  whose tail is of class  $\alpha$  and head is of class  $\beta$ , and similarly construct  $S$ -NNCT. All the asymptotic results in this chapter remain to hold for  $S$ -NNCT cell counts when  $r(d, k), q(d, k)$  and  $k$  are replaced with  $r(d, S), q(d, S)$  and  $\text{card}(S)$ , respectively.

## Chapter 4

## CLASSIFICATION OF ISOMORPHISM-INVARIANT RANDOM DIGRAPHS

## 4.1 Introduction

In this chapter, we classify isomorphism-invariant random digraphs according to where randomness resides, namely, arcs, vertices, and vertices and arcs together which in turn yield arc random digraphs (ARD), vertex random digraphs (VRD) and vertex-arc random digraphs (VARD), respectively. This digraph classification can be viewed as an extension of the classification of isomorphism-invariant random graphs. We introduce randomness in the direction of the edges of a given graph and obtain direction random digraphs (DRD) as well. We classify DRDs according to which component is random in addition to the direction and study the relations of DRDs with VARDs, VRDs and ARDs. We also consider random  $k$ NN digraphs and determine their membership with respect to these digraph families.

For a positive integer  $n$ , let  $[n] = \{1, 2, \dots, n\}$ ,  $\mathcal{D}_n$  denote the set of all digraphs with vertex set  $[n]$  and  $2^{\mathcal{D}_n}$  denote the set of all subsets of  $\mathcal{D}_n$ . A *random digraph* is a probability space  $(\mathcal{D}_n, 2^{\mathcal{D}_n}, P)$ , and we write  $\mathbf{D} = (\mathcal{D}_n, P)$  where  $P$  is a probability measure. We call a random digraph as *degenerate* if all the probability mass is on one digraph. We can also think of  $\mathbf{D}$  as the outcome of an experiment of picking a digraph from  $\mathcal{D}_n$  with distribution  $P$ . For every  $D \in \mathcal{D}_n$ , we write  $P(\{D\})$  as  $P(D)$  for brevity in notation. Also, for a measure space  $(\Omega, \mathcal{F}, \mu)$ ,  $\mathcal{F}^n$  and  $\mu^n$  denote the usual product  $\sigma$ -algebra and product measure, respectively. For the set of real numbers, we consider the Borel  $\sigma$ -algebra, and throughout this chapter we suppress the  $\sigma$ -algebra notation as long as there is no necessity nor ambiguity.

**Example 4.1.1.** (Uniform Random Digraph Model) For positive integers  $n$  and  $m$  with  $n \geq 2$  and  $0 < m < n(n-1)$ ,  $\mathbf{D}(n, m)$  is the random digraph such that

$$P(D) = \begin{cases} \frac{1}{\binom{n(n-1)}{m}}, & \text{if } |A(D)| = m \\ 0, & \text{otherwise} \end{cases}$$

for every  $D \in \mathcal{D}_n$ . In other words,  $\mathbf{D}(n, m)$  picks a digraph uniformly at random among the ones with vertex set  $[n]$  and having exactly  $m$  arcs. Note that there are  $\binom{n(n-1)}{m}$  such digraphs, and  $m$  is not chosen to be 0 or  $n(n-1)$  to obtain a non-degenerate random digraph. Also notice that  $\mathbf{D}(n, m)$  is the digraph version of the Erdős-Rényi random

graph  $\mathbf{G}(n, m)$  (Erdős and Rényi (1959)). For some asymptotic properties of uniform random digraphs see Luczak (1990) and Graham and Pike (2008).

**Definition 4.1.2.** (Isomorphism Invariance) Let  $\mathbf{D} = (\mathcal{D}_n, P)$  be a random digraph. We say that  $\mathbf{D}$  is *isomorphism-invariant* if  $P(D_1) = P(D_2)$  whenever  $D_1$  and  $D_2$  are isomorphic digraphs in  $\mathcal{D}_n$ .

Throughout the chapter, we only consider non-degenerate isomorphism-invariant random digraphs. We follow the isomorphism-invariant graph classification of Beer et al. (2011) pointing out the similarities and differences until we introduce randomness in the direction.

We next summarize isomorphism-invariant random graphs introduced by Beer et al. (2011). Let  $\mathcal{G}_n$  denote the set of all graphs with  $V(G) = [n]$  and  $2^{\mathcal{G}_n}$  be the set of all subsets of  $\mathcal{G}_n$ . A *random graph* is a probability space  $(\mathcal{G}_n, 2^{\mathcal{G}_n}, P)$ , and we write  $\mathbf{G} = (\mathcal{G}_n, P)$  where  $P$  is a probability measure.

The random graph model was first introduced by Gilbert (1959) and Erdős and Rényi (1959). The model of Gilbert corresponds to edge random graph  $\mathbf{G}(n, p_e)$  in Beer et al. (2011) in which each edge is inserted, independent of others, with probability  $p_e$ . The model introduced by Erdős and Rényi is the uniform random graph  $\mathbf{G}(n, m)$  which picks a graph with vertex set  $[n]$  uniformly at random among the ones with exactly  $m$  edges. However, in the literature, both of these models are usually called Erdős-Rényi model as they developed the theory.

We say that the random graph  $\mathbf{G} = (\mathcal{G}_n, P)$  is *isomorphism-invariant* if  $P(G_1) = P(G_2)$  whenever  $G_1$  is isomorphic to  $G_2$ .

**Definition 4.1.3.** An *edge random graph* (ERG) is a random graph  $\mathbf{G}(n, p_e) = (\mathcal{G}_n, P)$  where  $p_e \in [0, 1]$  and

$$P(G) = p_e^{|E(G)|} (1 - p_e)^{\binom{n}{2} - |E(G)|},$$

for every  $G \in \mathcal{G}_n$ .

Let  $\Omega$  be a set,  $\mathbf{x} = (x_1, \dots, x_n) \in \Omega^n$  and  $\phi : \Omega \times \Omega \rightarrow \{0, 1\}$  be a symmetric function. Then the  $(\mathbf{x}, \phi)$ -graph, denoted  $G(\mathbf{x}, \phi)$ , is defined to be the graph,  $G$ , with vertex set  $[n]$  such that for every  $i, j \in [n]$  with  $i \neq j$  we have  $ij \in E(G)$  if and only if  $\phi(x_i, x_j) = 1$ .

**Definition 4.1.4.** Let  $(\Omega, \mathcal{F}, \mu)$  be a probability space and  $\phi : \Omega \times \Omega \rightarrow \{0, 1\}$  be a symmetric measurable function. The *vertex random graph* (VRG),  $\mathbf{G}(n, \Omega, \mu, \phi)$ , is the random graph  $(\mathcal{G}_n, P)$  satisfying

$$P(G) = \int \mathbf{1}_{\{G(\mathbf{x}, \phi) = G\}} d(\mu \mathbf{x}) \text{ for every } G \in \mathcal{G}_n,$$

where  $d(\mu \mathbf{x})$  is short-hand for the product integrator  $d(\mu^n(\mathbf{x})) = d(\mu x_1) \cdots d(\mu x_n)$ .

Notice that in a VRG the randomness lies in the structure attached to the vertices, and once these random structures have been assigned to the vertices, all the edges are uniquely determined.

**Definition 4.1.5.** Let  $(\Omega, \mathcal{F}, \mu)$  be a probability space and  $\phi : \Omega \times \Omega \rightarrow [0, 1]$  be a symmetric measurable function. The *vertex-edge random graph* (VERG),  $\mathbf{G}(n, \Omega, \mu, \phi)$ , is the random graph  $(\mathcal{G}_n, P)$  with

$$P(G) = \int P_{\mathbf{x}}(G) d(\mu \mathbf{x}),$$

for every  $G \in \mathcal{G}_n$ , where for given  $\mathbf{x} = (x_1, \dots, x_n)$  and  $G$

$$P_{\mathbf{x}}(G) = \prod_{ij \in E(G)} \phi(x_i, x_j) \times \prod_{ij \notin E(G)} (1 - \phi(x_i, x_j)).$$

In words, a VERG is generated as follows: a random sample of size  $n$  is drawn with distribution  $\mu$  from  $\Omega$ , say  $\mathbf{X} = (X_1, \dots, X_n)$ . Then conditional on  $\mathbf{X}$ , independently for each pair of distinct vertices  $i$  and  $j$ , the edge  $ij$  is inserted with probability  $\phi(X_i, X_j)$ .

Observe that the same notation  $\mathbf{D}(n, \Omega, \mu, \phi)$  is used for both VRGs and VERGs. However, this causes no confusion, since  $\phi$  takes values in  $\{0, 1\}$  for VRGs and in  $[0, 1]$  for VERGs. In other words, VRGs form a special case of VERGs with  $\phi$  taking values only in  $\{0, 1\}$ . Therefore, every VRG is a VERG. In addition, it is easy to see that letting  $\phi$  to be identically equal to  $p$  gives that every ERG is a VERG.

Let  $\mathbf{G}_1 = (\mathcal{G}_n, P_1)$  and  $\mathbf{G}_2 = (\mathcal{G}_n, P_2)$  be random graphs. The *total variation distance* between  $\mathbf{G}_1$  and  $\mathbf{G}_2$  is defined to be

$$d_{\text{TV}}(\mathbf{G}_1, \mathbf{G}_2) = \frac{1}{2} \sum_{G \in \mathcal{G}_n} |P_1(G) - P_2(G)|.$$

Similarly, for any two random digraphs  $\mathbf{D}_1 = (\mathcal{D}_n, P_1)$  and  $\mathbf{D}_2 = (\mathcal{D}_n, P_2)$ , the *total variation distance* between  $\mathbf{D}_1$  and  $\mathbf{D}_2$  is defined to be

$$d_{\text{TV}}(\mathbf{D}_1, \mathbf{D}_2) = \frac{1}{2} \sum_{D \in \mathcal{D}_n} |P_1(D) - P_2(D)|.$$

## 4.2 ARDs, VRDs and VARDs

### 4.2.1 Arc Random Digraphs

One of the most commonly studied random digraphs is the binomial (or Bernoulli) random digraph model,  $\mathbf{D}(n, p_a)$ , in which each of the  $n(n-1)$  possible arcs is included independently with probability  $p_a$ . Such random digraphs give rise to arc random digraphs.

**Definition 4.2.1.** An *arc random digraph* (ARD) is a random digraph  $\mathbf{D}(n, p_a) = (\mathcal{D}_n, P)$  where  $0 < p_a < 1$  and

$$P(D) = p_a^{|A(D)|} (1 - p_a)^{n(n-1) - |A(D)|},$$

for every  $D \in \mathcal{D}_n$ .

Notice that ARDs are the digraph counterparts of random graphs  $\mathbf{G}(n, p_e)$  due to Gilbert (1959). For some asymptotic properties of  $\mathbf{D}(n, p_a)$  see Karp (1990), Luczak and Seierstad (2009), and Krivelevich et al. (2013).

**Definition 4.2.2.** Let  $p_a : [n] \times [n] \rightarrow [0, 1]$  be a function (that is not necessarily symmetric in its arguments). The *generalized arc random digraph* (GARD),  $\mathbf{D}(n, p_a)$ , is the random digraph  $(\mathcal{D}_n, P)$  with

$$P(D) = \prod_{(i,j) \in A(D)} p_a(i, j) \times \prod_{(i,j) \notin A(D)} (1 - p_a(i, j)),$$

for every  $D \in \mathcal{D}_n$ .

In other words, in a GARD each arc appears independently of others and the arc  $(i, j)$  occurs with probability  $p_a(i, j)$ . Note that an ARD is special case of a GARD with a constant  $p_a$ , i.e.,  $p_a(i, j) = p_a$  for all  $i, j$ . As the classical random digraph model  $\mathbf{D}(n, p_a)$  may not fit real life networks, inhomogeneous models like GARDs are of interest for such scenarios (see, e.g., Bloznelis et al. (2012)).

Clearly, any ARD is isomorphism-invariant. The following proposition implies that a GARD is isomorphism-invariant if and only if it is an ARD.

**Proposition 4.2.3.** *Let  $\mathbf{D}$  be an isomorphism-invariant GARD. Then  $\mathbf{D} = \mathbf{D}(n, p_a)$  for some  $p_a$ , i.e.,  $\mathbf{D}$  is an ARD.*

*Proof.* We show that  $p_a(i, j) = p_a(k, l)$  for any two ordered pairs  $(i, j)$  and  $(k, l)$ . First note that

$$p_a(i, j) = P((i, j) \in A(\mathbf{D})) = \sum_{(i,j) \in A(D)} P(D). \quad (4.1)$$

Fix a permutation on  $[n]$  which maps  $i$  to  $k$  and  $j$  to  $l$ . Observe that this permutation induces a one-to-one correspondence between the sets  $\{D \in \mathcal{D}_n : (i, j) \in A(D)\}$  and  $\{D' \in \mathcal{D}_n : (k, l) \in A(D')\}$  such that matched digraphs are isomorphic. As  $\mathbf{D}$  is isomorphism-invariant, this correspondence implies

$$\sum_{(i,j) \in A(D)} P(D) = \sum_{(k,l) \in A(D')} P(D'). \quad (4.2)$$

Hence, the result follows by (4.1) and (4.2).  $\square$

### 4.2.2 Vertex Random Digraphs

Let  $\Omega$  be a set,  $\mathbf{x} = (x_1, \dots, x_n) \in \Omega^n$  and  $\phi : \Omega \times \Omega \rightarrow \{0, 1\}$  be a function. Then the  $(\mathbf{x}, \phi)$ -digraph, denoted  $D(\mathbf{x}, \phi)$ , is defined to be the digraph,  $D$ , with vertex set  $[n]$  such that for all  $i, j \in [n]$  with  $i \neq j$  we have

$$(i, j) \in A(D) \text{ if and only if } \phi(x_i, x_j) = 1.$$

Clearly, every digraph  $D$  with  $V(D) = [n]$  is an  $(\mathbf{x}, \phi)$ -digraph for some choice of  $\Omega, \mathbf{x}$  and  $\phi$ . More specifically, choose  $\mathbf{x}$  to be the identity function on  $\Omega = [n]$  and define  $\phi(i, j) = \mathbf{1}_{\{(i, j) \in A(D)\}}$  where  $\mathbf{1}_{\{\cdot\}}$  is the indicator function.

**Definition 4.2.4.** Let  $(\Omega, \mathcal{F}, \mu)$  be a probability space and  $\phi : \Omega \times \Omega \rightarrow \{0, 1\}$  be a measurable function. The *vertex random digraph* (VRD),  $\mathbf{D}(n, \Omega, \mu, \phi)$ , is the random digraph  $(\mathcal{D}_n, P)$  with

$$P(D) = \int \mathbf{1}_{\{D(\mathbf{x}, \phi) = D\}} d(\mu \mathbf{x}),$$

for every  $D \in \mathcal{D}_n$ .

Note that in a VRD the randomness resides in the structure attached to the vertices, as in VRGs, and when these random structures are assigned to the vertices, all the arcs are uniquely determined.

**Example 4.2.5.** Proximity Catch Digraphs (PCDs) (Ceyhan (2011)): Let  $(\Omega, \mathcal{F}, \mu)$  be a probability space. The *proximity map*  $N(\cdot)$  is a function from  $\Omega$  to  $\mathcal{F}$ . The *proximity region* associated with  $x \in \Omega$ , denoted  $N(x)$ , is the image of  $x \in \Omega$  under  $N(\cdot)$ . The points in  $N(x)$  are thought of as being “closer” to  $x \in \Omega$  than the points in  $\Omega \setminus N(x)$ . For a given  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  the *proximity catch digraph* is the digraph with the vertex set  $V = [n]$  and the arc set  $A = \{(i, j) : x_j \in N(x_i)\}$ . In other words, we insert the arc  $(i, j)$  if and only if  $x_j$  is in the proximity region of  $x_i$ . Note that for a given  $N(\cdot)$ , a random PCD is a VRD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ , with  $\phi(x_i, x_j) = \mathbf{1}_{\{x_j \in N(x_i)\}}$ . For instance, one can take  $\Omega = \mathbb{R}$ ,  $N(x) = [x, \infty)$  and  $\phi(x, y) = \mathbf{1}_{\{x \leq y\}}$ .

**Example 4.2.6.** Random Intersection Digraphs (Bloznelis (2010)): Let  $n$  and  $m$  be positive integers, and  $\mu$  be a distribution on  $2^{[m]} \times 2^{[m]}$  (ordered pairs of subsets of  $[m]$ ). Given two collections of subsets  $S_1, \dots, S_n$  and  $T_1, \dots, T_n$  of the set  $[m]$ , define the intersection digraph with vertex set  $[n]$  such that the arc  $(i, j)$  is present in the digraph whenever  $S_i \cap T_j$  is nonempty for  $i \neq j$ .  $\mathbf{D}(n, m, \mu)$  is the random intersection digraph generated by independent and identically distributed pairs of random subsets  $(S_i, T_i)$  under  $\mu$ ,  $1 \leq i \leq n$ . Note that  $\mathbf{D}(n, m, \mu)$  is a VRD with  $\Omega = 2^{[m]} \times 2^{[m]}$  and  $\phi((S, T), (S', T')) = \mathbf{1}_{\{S \cap T' \neq \emptyset\}}$ .

By letting  $\Omega = [0, 1]$ ,  $\mu$  be the uniform distribution over  $[0, 1]$  and  $\phi(x, y) = \mathbf{1}_{\{x \leq p_a\}}$ , we see that every  $\mathbf{D}(2, p_a)$  is a VRD.

Recall that in a VRD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ ,  $\phi$  is not required to be symmetric. However, if  $\phi$  is a symmetric function, whenever we see the arc  $(i, j)$  in  $A(\mathbf{D})$ , we see the arc  $(j, i)$  as well. In this case, for every  $D \in \mathcal{D}_n$  in which there exists  $(i, j) \in A(D)$  with  $(j, i) \notin A(D)$ , we have  $P(D) = 0$ . On the other hand, in an ARD,  $\mathbf{D}(n, p_a)$ , we have  $P(D) > 0$  for every  $D \in \mathcal{D}_n$ . Therefore, whenever  $\phi$  is symmetric and nonconstant  $\mu^2$ -a.s.,  $\mathbf{D}(n, \Omega, \mu, \phi)$  is not an ARD. For instance, one can take  $\Omega = \mathbb{R}^d$ ,  $\mu$  to be an a.e. continuous distribution and  $\phi(x, y) = \mathbf{1}_{\{\|x-y\|_d \leq r\}}$ , where  $\|\cdot\|_d$  is the usual Euclidean norm in  $\mathbb{R}^d$  and  $r$  is a fixed positive real number. Notice that these random digraphs are random PCDs in which  $N(x)$  is the closed ball with radius  $r$  and center  $x$ . If we consider symmetric arcs as one edge only, these type of random digraphs reduce to what is called *random geometric graphs*. For more information about random geometric graphs see Penrose (2003).

#### 4.2.3 Vertex-Arc Random Digraphs

We now generalize the random digraphs introduced in the previous two subsections by combining the structures where the randomness lies.

**Definition 4.2.7.** Let  $(\Omega, \mathcal{F}, \mu)$  be a probability space and  $\phi : \Omega \times \Omega \rightarrow [0, 1]$  be a measurable function. The *vertex-arc random digraph* (VARD),  $\mathbf{D}(n, \Omega, \mu, \phi)$ , is the random digraph  $(\mathcal{D}_n, P)$  with

$$P(D) = \int P_{\mathbf{x}}(D) d(\mu \mathbf{x}),$$

for every  $D \in \mathcal{D}_n$ , where for given  $\mathbf{x} = (x_1, \dots, x_n)$  and  $D = (V, A)$

$$P_{\mathbf{x}}(D) = \prod_{(i,j) \in A} \phi(x_i, x_j) \times \prod_{(i,j) \notin A} (1 - \phi(x_i, x_j)).$$

The construction of a VARD is almost same with VERGs. A random sample of size  $n$  is drawn with distribution  $\mu$  from  $\Omega$ , say  $\mathbf{X} = (X_1, \dots, X_n)$ , and then conditional on  $\mathbf{X}$ , independently for each pair of distinct vertices  $i$  and  $j$ , the arc  $(i, j)$  is inserted with probability  $\phi(X_i, X_j)$ .

Note that we use the same notation  $\mathbf{D}(n, \Omega, \mu, \phi)$  for both VRDs and VARDs. But, since  $\phi$  takes values only 0 or 1 for VRDs and in  $[0, 1]$  for VARDs, this causes no confusion. Particularly, VRDs form a special case of VARDs with  $\phi$  taking values only in  $\{0, 1\}$ . Therefore, every VRD is a VARD. Moreover, it is easy to verify that letting  $\phi$  to be identically equal to  $p_a$  gives that every ARD is a VARD.

**Proposition 4.2.8.** *Every VARD is isomorphism-invariant.*

*Proof.* Let  $\mathbf{D}(n, \Omega, \mu, \phi)$  be a VARD and  $D, D' \in \mathcal{D}_n$  be isomorphic digraphs. Then there exists a permutation  $\sigma$  on  $[n]$  such that

$$(i, j) \in A(D) \Leftrightarrow (\sigma(i), \sigma(j)) \in A(D').$$

Let  $\sigma^{-1}$  be the inverse of  $\sigma$  and  $\mathbf{y} = (y_1, \dots, y_n)$  such that  $y_i = x_{\sigma^{-1}(i)}$  for all  $1 \leq i \leq n$ , i.e.,  $x_i = y_{\sigma(i)}$  for all  $1 \leq i \leq n$ . Then note that

$$\begin{aligned} P_{\mathbf{x}}(D) &= \prod_{(i,j) \in A(D)} \phi(x_i, x_j) \times \prod_{(i,j) \notin A(D)} (1 - \phi(x_i, x_j)) \\ &= \prod_{(i,j) \in A(D)} \phi(y_{\sigma(i)}, y_{\sigma(j)}) \times \prod_{(i,j) \notin A(D)} (1 - \phi(y_{\sigma(i)}, y_{\sigma(j)})) \\ &= \prod_{(\sigma(i), \sigma(j)) \in A(D')} \phi(y_{\sigma(i)}, y_{\sigma(j)}) \times \prod_{(\sigma(i), \sigma(j)) \notin A(D')} (1 - \phi(y_{\sigma(i)}, y_{\sigma(j)})) \\ &= \prod_{(i,j) \in A(D')} \phi(y_i, y_j) \times \prod_{(i,j) \notin A(D')} (1 - \phi(y_i, y_j)) \\ &= P_{\mathbf{y}}(D'). \end{aligned} \tag{4.3}$$

As  $\mathbf{y}$  is a permutation of  $\mathbf{x}$ , Fubini's theorem and (4.3) imply that

$$P(D) = \int P_{\mathbf{x}}(D) \mu(d\mathbf{x}) = \int P_{\mathbf{y}}(D') \mu(d\mathbf{y}). \tag{4.4}$$

Furthermore, the change of variables that maps  $y_i$  to  $x_i$  in the integrant above results

$$\int P_{\mathbf{y}}(D') \mu(d\mathbf{y}) = \int P_{\mathbf{x}}(D') \mu(d\mathbf{x}) = P(D'), \tag{4.5}$$

since the mapping is a permutation and the Jacobian of a permutation matrix is  $\pm 1$ . Thus, the results in (4.4) and (4.5) together imply that  $P(D) = P(D')$ , and so the desired result follows.  $\square$

As a corollary, we easily see that any VRD is isomorphism-invariant since every VRD is a VARD.

### 4.3 Inclusion/Exclusion Relations between ARDs, VRDs and VARDs

In the previous section we have shown that every ARD is a VARD and so is every VRD, and every VARD is isomorphism-invariant. In this section we prove that for  $n \geq 4$  there exists no random digraph which is both ARD and VRD, and the union of the classes ARDs and VRDs is not the entire class of VARDs.

The following theorem implies that the families ARDs and VRDs are disjoint for  $n \geq 4$ .

**Theorem 4.3.1.** *If an ARD,  $\mathbf{D}(n, p_a)$ , with  $n \geq 4$  is represented as a VARD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ , then  $\phi(x, y) = p_a \mu^2$ -a.s.*

*Proof.* Suppose that an ARD,  $\mathbf{D}(n, p_a)$ , with  $n \geq 4$  is represented as a VARD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ . For the proof of the theorem, we borrow some tools from functional analysis which are presented in the proof of the Theorem 4.2. in Beer et al. (2011). Let  $h : \Omega \times \Omega \rightarrow [0, 1]$  be a symmetric measurable function and  $T$  be the integral operator with kernel  $h$  on the space  $L^2(\Omega, \mu)$  of  $\mu$ -square-integrable functions on  $\Omega$ :

$$(Tg)(x) = \int h(x, y)g(y)d(\mu y).$$

Since  $h$  is bounded and  $\mu$  is a finite measure, the kernel  $h$  is in  $L^2(\mu \times \mu)$ . Integral operators with such kernels are Hilbert-Schmidt operators and are thus compact operators. Moreover, as  $h$  is symmetric, the integral operator  $T$  is self-adjoint, which implies that  $L^2(\Omega, \mu)$  has an orthonormal basis  $(\psi_i)_{i \geq 1}$  of eigenfunctions for  $T$  such that  $T\psi_i = \lambda_i \psi_i$  for not necessarily distinct real eigenvalues  $\lambda_i$  with  $\lambda_i \rightarrow 0$  as  $i \rightarrow \infty$  (see Chapter VI in Reed and Simon (1980)). We may assume that  $\lambda_1$  is the largest eigenvalue. Then we have

$$h(x, y) = \sum_{i \geq 1} \lambda_i \psi_i(x) \psi_i(y) \quad \mu^2\text{-a.s.}$$

with the sum converging in  $L^2$ . As  $\psi_i$ 's are orthonormal, it follows that

$$\begin{aligned} & \mathbf{E}(h(X_1, X_2)h(X_2, X_3)h(X_3, X_4)h(X_4, X_1)) \\ &= \int \int \int \int h(x_1, x_2)h(x_2, x_3)h(x_3, x_4)h(x_4, x_1)d(\mu x_1)d(\mu x_2)d(\mu x_3)d(\mu x_4) \\ &= \sum_{i \geq 1} \lambda_i^4. \end{aligned} \tag{4.6}$$

Now let  $E_1$  be the event that both  $(1, 2)$  and  $(2, 1)$  are in  $A(\mathbf{D})$ . As  $\mathbf{D}$  is an ARD  $\mathbf{D}(n, p_a)$ , it is easy to see that  $P(E_1) = p_a^2$ . On the other hand, since  $\mathbf{D}$  is represented as a VARD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ , we have

$$P(E_1) = \mathbf{E}(\phi(X_1, X_2)\phi(X_2, X_1)).$$

Thus, letting  $h(x, y) = \phi(x, y)\phi(y, x)$  gives  $p_a^2 = \mathbf{E}(h(X_1, X_2))$ . As

$$\mathbf{E}(h(X_1, X_2)) = \int \int h(x, y)d(\mu x)d(\mu y) = \langle T\mathbf{1}, \mathbf{1} \rangle \leq \lambda_1,$$

we get  $p_a^2 \leq \lambda_1$ , where  $\mathbf{1}$  is the function with constant value 1.

Let  $E_2$  be the event that  $(1, 2), (2, 1), (2, 3), (3, 2), (3, 4), (4, 3), (4, 1), (1, 4) \in A(\mathbf{D})$ .

Since  $\mathbf{D}$  is an ARD,  $\mathbf{D}(n, p_a)$ , it is easy to see that

$$P(E_2) = p_a^8. \quad (4.7)$$

By the representation of  $\mathbf{D}$  as a VARD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ , we also have

$$P(E_2) = \mathbf{E}(h(X_1, X_2)h(X_2, X_3)h(X_3, X_4)h(X_4, X_1)). \quad (4.8)$$

Now combining the results in (4.6), (4.7) and (4.8) gives

$$p_a^8 = \sum_{i \geq 1} \lambda_i^4. \quad (4.9)$$

Since  $p_a^2 \leq \lambda_1$ , we have  $p_a^8 \leq \lambda_1^4$  and thus, by (4.9) we obtain that  $\lambda_1 = p_a^2$  and  $\lambda_i = 0$  for every  $i \geq 2$ , that is  $h(x, y) = p_a^2 \psi_1(x) \psi_1(y)$ . But then we have

$$\begin{aligned} p_a^2 \int \psi_1^2(x) d(\mu x) &= p_a^2 = \mathbf{E}(h(X_1, X_2)) \\ &= p_a^2 \int \int \psi_1(x) \psi_1(y) d(\mu x) d(\mu y) = p_a^2 \left( \int \psi_1(x) d(\mu x) \right)^2, \end{aligned}$$

which implies that

$$\int \psi_1^2(x) d(\mu x) = \left( \int \psi_1(x) d(\mu x) \right)^2, \quad (4.10)$$

since  $p \neq 0$ . As the equality in equation (4.10) is the equality in the Cauchy-Schwarz inequality for  $\psi_1$  and  $\mathbf{1}$ , we see that  $\psi_1$  is constant  $\mu$ -a.s. Since  $\int \psi_1^2(x) d(\mu x) = 1$ , we get  $\psi_1 = 1$   $\mu$ -a.s. or  $\psi_1 = -1$   $\mu$ -a.s., and therefore  $h(x, y) = p_a^2 \mu^2$ -a.s., that is

$$\phi(x, y) \phi(y, x) = p_a^2 \quad \mu^2\text{-a.s.} \quad (4.11)$$

Next, let  $E_3$  be the event that neither of the arcs  $(1, 2)$  and  $(2, 1)$  is in  $A(\mathbf{D})$ , and  $E_4$  be the event that none of the arcs  $(1, 2), (2, 1), (2, 3), (3, 2), (3, 4), (4, 3), (4, 1), (1, 4)$  is in  $A(\mathbf{D})$ . Choosing  $h(x, y)$  to be  $(1 - \phi(x, y))(1 - \phi(y, x))$  allows us to follow the same arguments above for  $E_1$  and  $E_2$  replaced with  $E_3$  and  $E_4$ , respectively, and with  $1 - p_a$  taking place of  $p_a$ . Therefore, we obtain that

$$(1 - \phi(x, y))(1 - \phi(y, x)) = (1 - p_a)^2 \quad \mu^2\text{-a.s.} \quad (4.12)$$

Finally, the equations in (4.11) and (4.12) give the desired result.  $\square$

*Remark 4.3.2.* The function  $h$  in the proof of Theorem 4.3.1 is taken to be symmetric. Otherwise, the operator  $T$  needs not to be self-adjoint and hence the succeeding arguments in the proof are not true. So,  $h$  being symmetric is a crucial condition for the

proof. If  $\phi$  is given to be symmetric, one can take  $h = \phi$  and obtain  $\phi = p_a$   $\mu^2$ -a.s. (as in the proof of Theorem 4.2. in Beer et al. (2011)). However, in our case,  $\phi$  is not supposed to be symmetric. Note that we tackle this hurdle by taking  $h(x, y)$  to be  $\phi(x, y)\phi(y, x)$  and  $(1 - \phi(x, y))(1 - \phi(y, x))$ , respectively, and obtain the desired result.

As any VRD is a VARD with  $\phi$  taking values in  $\{0, 1\}$ , by Theorem 4.3.1 we have the following corollary.

**Corollary 4.3.3.** *Any ARD,  $\mathbf{D}(n, p_a)$ , with  $n \geq 4$  is not a VRD.*

We next show that union of ARD and VRD families do not constitute the entire class of VARDs when  $n \geq 4$ .

**Theorem 4.3.4.** *There exist VARDs with  $n \geq 4$  which are neither a VRD nor an ARD.*

*Proof.* Let  $0 < a < b < 1$  be real numbers. Consider a VARD,  $\mathbf{D}(n, \Omega, \mu, \psi)$ , with  $n \geq 4$  such that  $\phi(x, y) \in \{a, b\}$ , and  $\psi(x, y) \neq \psi(y, x)$  for any  $x \neq y$ . Equivalently, we have

$$\psi(x, y)\psi(y, x) = ab \text{ and } (1 - \psi(x, y))(1 - \psi(y, x)) = (1 - a)(1 - b) \text{ for } x \neq y. \quad (4.13)$$

For example, one can take  $\Omega = \mathbb{R}$ ,  $\mu$  to be a continuous distribution and  $\psi(x, y) = a\mathbf{1}_{\{x \leq y\}} + b\mathbf{1}_{\{y < x\}}$ .

Now suppose that it has another VARD representation  $\mathbf{D}(n, \Omega', \nu, \phi)$ . We claim that  $\phi$  satisfies the same properties of  $\psi$  given in (4.13)  $\nu^2$ -a.s. Recall that in the proof of Theorem 4.3.1, the properties of an ARD,  $\mathbf{D}(n, p_a)$ , that we used are

$$P(E_2) = p_a^8 = (p_a^2)^4 = (P(E_1))^4 \quad (4.14)$$

and

$$P(E_4) = (1 - p_a)^8 = ((1 - p_a)^2)^4 = (P(E_3))^4. \quad (4.15)$$

Notice that the equations in (4.14) and (4.15) hold for  $\mathbf{D}(n, \Omega, \mu, \psi)$  when  $p_a^2$  and  $(1 - p_a)^2$  are replaced with  $ab$  and  $(1 - a)(1 - b)$ , respectively. That is, for  $\mathbf{D}(n, \Omega, \mu, \psi)$  we have

$$P(E_2) = (ab)^4 = (P(E_1))^4 \text{ and } P(E_4) = ((1 - a)(1 - b))^4 = (P(E_3))^4.$$

Therefore, following the same arguments in the proof of Theorem 4.3.1 we obtain

$$\phi(x, y)\phi(y, x) = ab \text{ and } (1 - \phi(x, y))(1 - \phi(y, x)) = (1 - a)(1 - b) \quad \nu^2\text{-a.s.}$$

and hence the claim follows. Then, by the choice of  $a, b$  and  $\psi$ , we see that  $\mathbf{D}(n, \Omega, \mu, \psi)$  has neither an ARD nor a VRD representation.  $\square$

Recall that Corollary 4.3.3 and Theorem 4.3.4 imply that for  $n \geq 4$   $\text{ARD} \cap \text{VRD} = \emptyset$  and  $\text{VARD} \setminus (\text{ARD} \cup \text{VRD}) \neq \emptyset$ , respectively, in Figure 4.1.

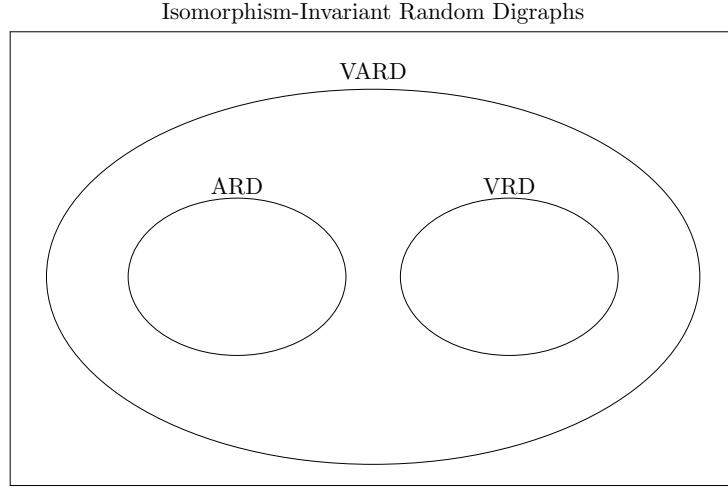


Figure 4.1: Venn diagram of vertex-arc random digraphs for  $n \geq 4$ . The results of this paper show that  $\text{ARD} \cap \text{VRD} = \emptyset$  and all the four regions in the figure are nonempty for  $n \geq 4$ .

**Remark 4.3.5. Approximation to VARDs by VRDs:** However, any VARD can be arbitrarily closely approximated by VRDs. That is, for any VARD  $\mathbf{D}$  and  $\epsilon > 0$ , there exists a VRD  $\mathbf{D}'$  such that  $d_{\text{TV}}(\mathbf{D}, \mathbf{D}') < \epsilon$ . This result is a straightforward extension of approximation of VERGs by VRGs which immediately follows by letting  $\psi(y_1, y_2)$  to be the indicator of the event  $\phi(x_1, x_2) \geq f_1(a_2)$  in the proof of Theorem 3.3 in Beer et al. (2011).

#### 4.4 Direction Random Digraphs

One can also obtain isomorphism-invariant random digraphs by first generating an isomorphism-invariant random graph and then assigning directions randomly to each edge. Along this line, we first generate an isomorphism-invariant random graph,  $\mathbf{G} = (\mathcal{G}_n, P_{\mathbf{G}})$ , and then for each edge  $ij \in E(\mathbf{G})$ , independent of other edges, pick a one sided or two sided direction randomly between  $i$  and  $j$ . For a given direction probability  $1/2 \leq p_d < 1$ , we put only the arc  $(i, j)$  with probability  $1 - p_d$ , only the arc  $(j, i)$  with probability  $1 - p_d$  and both of the arcs with probability  $2p_d - 1$ . Observe that the arc  $(i, j)$  is put with probability  $p_d$ . Also, note that we omit the case  $p_d = 1$  because it removes randomness in the direction.

The *underlying graph* of a digraph  $D$ , denoted  $U(D)$ , is the graph obtained by replacing each arc of  $D$  with an edge, disallowing multiple edges between two vertices (Chartrand and Lesniak (1996)).

**Definition 4.4.1.** The *underlying random graph* of a random digraph  $\mathbf{D} = (\mathcal{D}_n, P_{\mathbf{D}})$  is

the random graph  $\mathbf{G} = (\mathcal{G}_n, P_{\mathbf{G}})$  such that

$$P_{\mathbf{G}}(G) = \sum_{U(D)=G} P_{\mathbf{D}}(D),$$

for every  $G \in \mathcal{G}_n$ .

For instance, the underlying random graph of an ARD,  $\mathbf{D}(n, p_a)$ , is an ERG, namely,  $\mathbf{G}(n, p_e)$  with  $p_e = 2p_a - p_a^2$ . Moreover, notice also that the underlying random graph of a VARD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ , is the VERG,  $\mathbf{G}(n, \Omega, \mu, \phi_u)$ , where  $\phi_u(x, y) = \phi(x, y) + \phi(y, x) - \phi(x, y)\phi(y, x)$ . In particular, the underlying random graph of a VRD is a VRG.

For a digraph  $D \in \mathcal{D}_n$ , let  $n_a(D) = |A(D)|$  and  $n_e(D) = |E(U(D))|$  (i.e., the number of edges of the underlying graph of  $D$ ). Also, let  $n_s(D)$  denote the number of pairs of vertices  $i$  and  $j$  such that both  $(i, j)$  and  $(j, i)$  are in  $A(D)$  (i.e., the number of symmetric arcs in  $D$ ), and  $n_{as}(D)$  denote the number of arcs  $(i, j)$  in  $A(D)$  with  $(j, i) \notin A(D)$ . We write  $n_a, n_e, n_s$  and  $n_{as}$ , respectively, dropping the digraph  $D$  in the notation for brevity. Note that  $n_e = n_s + n_{as}$  and  $n_a = 2n_s + n_{as}$ .

**Definition 4.4.2.** Let  $\mathbf{G} = (\mathcal{G}_n, P_{\mathbf{G}})$  be an isomorphism-invariant random graph and  $1/2 \leq p_d < 1$ . A *direction random digraph* (DRD) is a random digraph  $\mathbf{D} = (\mathcal{D}_n, P)$  with

$$P(D) = P_{\mathbf{G}}(U(D))(1 - p_d)^{n_{as}}(2p_d - 1)^{n_s},$$

for every  $D \in \mathcal{D}_n$ , and we say that  $\mathbf{D}$  is *generated by*  $\mathbf{G}$  with *direction probability*  $p_d$ .

A natural question is why not start with a non-random graph and insert directions randomly to the edges to obtain DRDs. There is a simple answer to the question. Unfortunately, if directions are randomly inserted to the edges of a (fixed) graph, the resulting random digraph is not isomorphism-invariant unless we start with an *empty graph* (the graph with no edges) or a *complete graph* (the graph with all possible edges). Notice that  $\mathbf{G}$  is the underlying random graph of a DRD generated by  $\mathbf{G}$ . Observe that if the digraphs  $D_1$  and  $D_2$  are isomorphic, then so are the (underlying) graphs  $U(D_1)$  and  $U(D_2)$ , and we also have  $n_s(D_1) = n_s(D_2)$  and  $n_{as}(D_1) = n_{as}(D_2)$ . Thus, a DRD is isomorphism-invariant only if it is generated by an isomorphism-invariant random graph. moreover, notice that we may consider a (fixed) graph as a degenerate random graph. Also, it is easy to see that the empty graph and the complete graph with vertex set  $[n]$  are the only graphs in  $\mathcal{G}_n$  which are isomorphic to no other graph in  $\mathcal{G}_n$ , and therefore these two graphs are the only isomorphism-invariant degenerate random graphs.

#### 4.4.1 DERDs, DVRDs and DVERDs

We next provide three classes of direction random digraphs which are generated by ERGs, VRGs or VERGs.

**Definition 4.4.3.** The direction random digraph generated by an ERG,  $\mathbf{G}(n, p_e)$ , with direction probability  $p_d$  is called *direction-edge random digraph* (DERD) and denoted  $\mathbf{D}(n, p_e, p_d)$ .

Notice that letting  $p_d$  to be  $1/2$  avoids symmetric arcs, and hence in the case of  $p_d = 1/2$ , after generating an ERG each edge is independently oriented in one of the two directions with equal probability (e.g., see the model in Subramanian (2003)). For example, letting  $p_e = 1$  and  $p_d = 1/2$  gives a *random tournament* in which each edge of a complete graph is independently oriented in one direction with equal probability. For more information about tournaments, see Moon (1968).

**Definition 4.4.4.** A direction random digraph generated by a VRG is called *direction-vertex random digraph* (DVRD). A *direction-vertex-edge random digraph* (DVERD) is a direction random digraph generated by a VERG.

Notice that the underlying random graphs of a DERD, a DVRD and a DVERD are an ERG, a VRG and a VERG, respectively. Clearly any ERG or VRG is a VERG, and hence every DERD and DVRD has a DVERD representation. In addition, the results in Beer et al. (2011) imply the following: A non-degenerate DRD which is both a DERD and a DVRD is either with  $n \leq 3$  or generated by an ERG,  $\mathbf{G}(n, p_e)$ , with  $p_e = 1$ . For every  $n \geq 6$ , there exist DVERDs which are neither DERDs nor DVRDs. Moreover, for  $n \geq 3$ , there exist DRDs which are not among DVERDs, and for  $n \leq 3$ , any DVERD is also a DVRD. These results are illustrated in Figure 4.2.

**Remark 4.4.5. Approximation to DVERDs by DVRDs:** Let  $\mathbf{D} = (\mathcal{D}_n, P_{\mathbf{D}})$  be a DVERD generated by a VERG,  $\mathbf{G} = (\mathcal{G}_n, P_{\mathbf{G}})$ , with direction probability  $p_d$ . By Theorem 3.3 in Beer et al. (2011), for any  $\epsilon > 0$  there exists a VRG,  $\mathbf{G}' = (\mathcal{G}_n, P_{\mathbf{G}'})$ , satisfying  $d_{\text{TV}}(\mathbf{G}, \mathbf{G}') < \epsilon$ . Let  $\mathbf{D}' = (\mathcal{D}_n, P_{\mathbf{D}'})$  be the DVRD generated by  $\mathbf{G}'$  with the same direction probability  $p_d$ . Then, it is easy to see that

$$\sum_{U(D)=G} |P_{\mathbf{D}}(D) - P_{\mathbf{D}'}(D)| = |P_{\mathbf{G}}(G) - P_{\mathbf{G}'}(G)|$$

for every  $G \in \mathcal{G}_n$ , and therefore we get  $d_{\text{TV}}(\mathbf{D}, \mathbf{D}') = d_{\text{TV}}(\mathbf{G}, \mathbf{G}')$  which implies that the total deviation distance between  $\mathbf{D}$  and  $\mathbf{D}'$  is less than  $\epsilon$ .

#### 4.5 Inclusion/Exclusion Relations of DERDs with respect to VARs

In this section, for  $n \geq 4$ , we show that a random digraph is both a DERD and a VARD if and only if it is an ARD, and any DERD with  $n \leq 3$  is also a VARD.

**Proposition 4.5.1.** A DERD,  $\mathbf{D}(n, p_e, p_d)$ , is an ARD,  $\mathbf{D}(n, p_a)$ , if and only if

$$p_d = \frac{1}{1 + \sqrt{1 - p_e}} \text{ and } p_a = 1 - \sqrt{1 - p_e}.$$

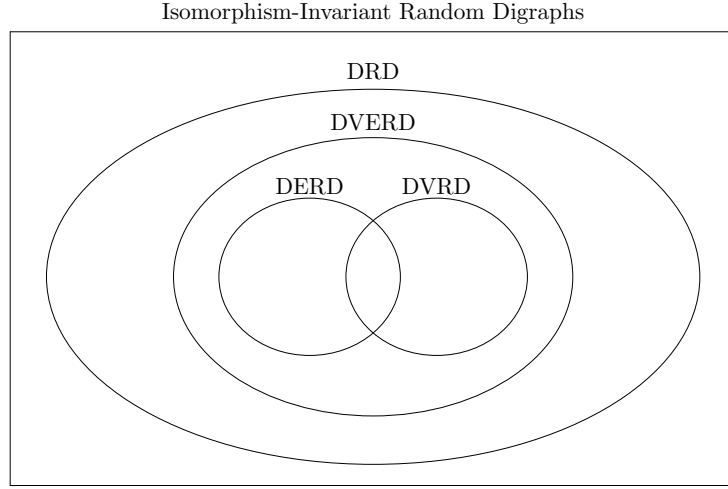


Figure 4.2: Venn diagram of direction random digraphs. The results of the paper imply that all the six regions in the figure are nonempty. In particular, the region  $\text{DERD} \cap \text{DVRD}$  only consists of DRDs with  $n \leq 3$  and DRDs generated by  $\mathbf{G}(n, p_e = 1)$ .

*Proof.* Suppose that  $\mathbf{D}(n, p_e, p_d)$  is an ARD  $\mathbf{D}(n, p_a)$ . Then we have  $p_e p_d = p_a$  since both are  $P((1, 2) \in A(\mathbf{D}))$ . Similarly we have  $p_e(2p_d - 1) = p_a^2$  as both are  $P(\{(1, 2), (2, 1)\} \subset A(\mathbf{D}))$ . Solving these two equations gives  $p_d = (1 \pm \sqrt{1 - p_e})/p_e$ . If  $p_e = 1$ , then definitely  $p_d = 1$ . Otherwise,  $(1 - \sqrt{1 - p_e})/p_e < 1 < (1 + \sqrt{1 - p_e})/p_e$  and hence  $p_d = (1 - \sqrt{1 - p_e})/p_e$ . Note that  $(1 - \sqrt{1 - p_e})/p_e = 1/(1 + \sqrt{1 - p_e})$  and so  $1/2 \leq p_d \leq 1$ . Finally,  $p_a = p_e p_d = 1 - \sqrt{1 - p_e}$ .

We next show that whenever  $p_d = 1/(1 + \sqrt{1 - p_e})$  and  $p_a = 1 - \sqrt{1 - p_e}$ ,  $\mathbf{D}(n, p_e, p_d)$  is  $\mathbf{D}(n, p_a)$ . Note that in that case,  $1 - p_d = p_a(1 - p_a)/p_e$ ,  $2p_d - 1 = p_a^2/p_e$  and  $1 - p_e = (1 - p_a)^2$ . Therefore, for a given  $D \in \mathcal{D}_n$  we have

$$\begin{aligned}
 P(D) &= p_e^{n_e} (1 - p_e)^{\binom{n}{2} - n_e} (1 - p_d)^{n_{as}} (2p_d - 1)^{n_s} \\
 &= p_e^{n_e} (1 - p_e)^{(n(n-1)-2n_e)} \frac{p_a^{n_{as}} (1 - p_a)^{n_{as}}}{p_e^{n_{as}}} \frac{p_a^{2n_s}}{p_e^{n_s}} \\
 &= p_e^{n_e - n_{as} - n_s} p_a^{n_{as} + 2n_s} (1 - p_a)^{n(n-1) - 2n_e + n_{as}} \\
 &= p_a^{n_a} (1 - p_a)^{n(n-1) - n_a},
 \end{aligned}$$

since  $n_e = n_{as} + n_s$  and  $n_a = n_{as} + 2n_s$ . Thus, the desired result follows.  $\square$

In fact, for  $n \geq 4$ , the family of ARDs is the intersection of the classes DERDs and VARDs.

**Theorem 4.5.2.** *If a DERD,  $\mathbf{D}(n, p_e, p_d)$ , with  $n \geq 4$  has a VARD representation  $\mathbf{D}(n, \Omega, \mu, \phi)$ , then  $p_d = 1/(1 + \sqrt{1 - p_e})$  and  $\phi(x, y) = p_e p_d \mu^2$ -a.s.*

*Proof.* Suppose  $\mathbf{D}(n, p_e, p_d)$  has a VARD representation  $\mathbf{D}(n, \Omega, \mu, \phi)$ . Note that, as in any ARD, the events  $(i, j) \in A(\mathbf{D})$  and  $(k, l) \in A(\mathbf{D})$  are independent in a DERD whenever  $\{i, j\} \neq \{k, l\}$ . Therefore, one can apply the method used in the proof of Theorem 4.3.1 and obtain

$$\phi(x, y)\phi(y, x) = p_e(2p_d - 1) \quad \mu^2\text{-a.s.} \quad (4.16)$$

and

$$(1 - \phi(x, y))(1 - \phi(y, x)) = 1 - p_e \quad \mu^2\text{-a.s.} \quad (4.17)$$

Solving the equations in (4.16) and (4.17) yields

$$\phi(x, y) + \phi(y, x) = 2p_e p_d \quad \mu^2\text{-a.s.} \quad (4.18)$$

and

$$\phi(x, y) = p_e p_d \pm \sqrt{(1 - \sqrt{1 - p_e} - p_e p_d)(1 + \sqrt{1 - p_e} - p_e p_d)} \quad \mu^2\text{-a.s.} \quad (4.19)$$

If  $p_d > 1/(1 + \sqrt{1 - p_e}) = (1 - \sqrt{1 - p_e})/p_e$ , then the numbers in the right-hand side of (4.19) have imaginary parts, and hence we get a contradiction since  $\phi$  takes only real values.

Suppose  $p_d \leq 1/(1 + \sqrt{1 - p_e})$ . In any VARD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ , with  $n \geq 3$  we have

$$\begin{aligned} P(\{(1, 2), (1, 3)\} \subset A(\mathbf{D})) &= \int \int \int \phi(x_1, x_2)\phi(x_1, x_3)d(\mu x_1)d(\mu x_2)d(\mu x_3) \\ &= \int \left( \int \phi(x_1, x_2)d(\mu x_2) \right) \left( \int \phi(x_1, x_3)d(\mu x_3) \right) d(\mu x_1) \\ &= \int \left( \int \phi(x_1, x_2)d(\mu x_2) \right)^2 d(\mu x_1) \\ &\geq \left( \int \int \phi(x_1, x_2)d(\mu x_1)d(\mu x_2) \right)^2 \\ &= (P((1, 2) \in A(\mathbf{D})))^2 \end{aligned} \quad (4.20)$$

by Fubini's theorem and the Cauchy-Schwarz inequality applied to the constant function 1 and  $\int \phi(x_1, x_2)d(\mu x_2)$ . On the other hand, in any DERD,  $\mathbf{D}(n, p_e, p_d)$ , with  $n \geq 3$  we have

$$P(\{(1, 2), (1, 3)\} \subset A(\mathbf{D})) = (p_e p_d)^2 = (P((1, 2) \in A(\mathbf{D})))^2.$$

Therefore, we have the equality in the Cauchy-Schwarz inequality in (4.20). Thus,

$\int \phi(x_1, x_2) d(\mu x_2) = c$   $\mu$ -a.s. for some constant  $c$ . Since

$$p_e p_d = P((1, 2) \in A(\mathbf{D})) = \int \int \phi(x_1, x_2) d(\mu x_2) d(\mu x_1) = \int c d(\mu x_1) = c,$$

we obtain  $c = p_e p_d$ , that is,

$$\int \phi(x, y) d(\mu y) = p_e p_d \quad \mu\text{-a.s.} \quad (4.21)$$

Similarly, in a VARD,  $\mathbf{D}(n, \Omega, \mu, \phi)$ , with  $n \geq 4$  we have

$$\begin{aligned} & P(\{(1, 2), (2, 3), (1, 4), (4, 3)\} \subset A(\mathbf{D})) \\ &= \int \int \int \int \phi(x_1, x_2) \phi(x_2, x_3) \phi(x_1, x_4) \phi(x_4, x_3) d(\mu x_1) d(\mu x_2) d(\mu x_3) d(\mu x_4) \\ &= \int \int \left( \int \phi(x_1, x_2) \phi(x_2, x_3) d(\mu x_2) \right) \left( \int \phi(x_1, x_4) \phi(x_4, x_3) d(\mu x_4) \right) d(\mu x_1) d(\mu x_3) \\ &= \int \int \left( \int \phi(x_1, x_2) \phi(x_2, x_3) d(\mu x_2) \right)^2 d(\mu x_1) d(\mu x_3) \\ &\geq \left( \int \int \int \phi(x_1, x_2) \phi(x_2, x_3) d(\mu x_1) d(\mu x_2) d(\mu x_3) \right)^2 \\ &= (P(\{(1, 2), (2, 3)\} \subset A(\mathbf{D})))^2 \end{aligned} \quad (4.22)$$

by Fubini's theorem and the Cauchy-Schwarz inequality applied to the constant function **1** and  $\int \phi(x_1, x_2) \phi(x_2, x_3) d(\mu x_2)$ . Since in a DERD,  $\mathbf{D}(n, p_e, p_d)$ , with  $n \geq 4$  we have

$$P(\{(1, 2), (2, 3), (1, 4), (4, 3)\} \subset A(\mathbf{D})) = (p_e p_d)^4 = (P(\{(1, 2), (2, 3)\} \subset A(\mathbf{D})))^2,$$

we obtain  $\int \phi(x_1, x_2) \phi(x_2, x_3) d(\mu x_2)$  is constant  $\mu^2$ -a.s. by the equality in Cauchy-Schwarz inequality in (4.22). By the equality in (4.22), one can easily verify that

$$\int \phi(x, y) \phi(y, z) d(\mu y) = (p_e p_d)^2 \quad \mu^2\text{-a.s.} \quad (4.23)$$

Let  $s(x, y) = i(\phi(x, y) - p_e p_d)$ . Combining the results in (4.18), (4.19), (4.21) and (4.23) gives

$$s(x, y) = i(\phi(x, y) - p_e p_d) = -i(\phi(y, x) - p_e p_d) = \overline{s(y, x)} \quad \mu^2\text{-a.s.} \quad (4.24)$$

and

$$\int s(x, y) s(y, z) d(\mu y) = 0 \quad \mu^2\text{-a.s.} \quad (4.25)$$

Let  $T$  be the integral operator with kernel  $s$  on the space  $L^2(\Omega, \mu)$

$$(Tg)(x) = \int s(x, y)g(y)d(\mu y).$$

Since  $s$  is bounded and  $\mu$  is a finite measure, the kernel  $s$  is in  $L^2(\mu \times \mu)$ . Moreover, the integral operator  $T$  is compact and self-adjoint by (4.24), which implies that  $L^2(\Omega, \mu)$  has an orthonormal basis  $(\psi_i)_{i \geq 1}$  of eigenfunctions for  $T$  such that  $T\psi_i = \lambda_i\psi_i$  for not necessarily distinct eigenvalues  $\lambda_i$ , and

$$s(x, y) = \sum_{i \geq 1} \lambda_i \psi_i(x) \overline{\psi_i(y)} \quad \mu^2\text{-a.s.} \quad (4.26)$$

with the sum converging in  $L^2$  (Reed and Simon (1980)). Since  $\psi_i$ 's are orthonormal, the equations (4.25) and (4.26) imply

$$\sum_{i \geq 1} \lambda_i^2 \psi_i(x) \overline{\psi_i(z)} = 0 \quad \mu^2\text{-a.s.} \quad (4.27)$$

Therefore, for any  $m \geq 1$ , by multiplying the equation in (4.27) by  $\psi_m(z) \overline{\psi_m(x)}$  and integrating over  $x$  and  $z$  we obtain  $\lambda_m^2 = 0$ , i.e.,  $\lambda_m = 0$  for each  $m$ . Thus,  $s(x, y) = 0$   $\mu^2$ -a.s. and hence  $\phi(x, y) = p_e p_d$   $\mu^2$ -a.s. which implies  $p_d = 1/(1 + \sqrt{1 - p_e})$ .  $\square$

*Remark 4.5.3.* Notice that Theorem 4.5.2 and Proposition 4.5.1 together imply Theorem 4.3.1. However, we provide the proof of Theorem 4.3.1 to keep the proof of Theorem 4.5.2 shorter and also to point out similarities and the differences with the techniques used in Beer et al. (2011).

However, for  $n \leq 3$ , any DERD has a VARD representation.

**Theorem 4.5.4.** *Any DERD,  $\mathbf{D}(n, p_e, p_d)$ , with  $n \leq 3$  is also a VARD.*

*Proof.* Let  $\oplus$  and  $\ominus$  denote addition and subtraction modulo 1, respectively. In other words, for real numbers  $0 \leq x, y < 1$ ,

$$x \oplus y = \begin{cases} x + y, & \text{if } x + y < 1 \\ x + y - 1, & \text{if } x + y \geq 1 \end{cases}$$

and

$$x \ominus y = \begin{cases} x - y, & \text{if } x - y \geq 0 \\ x - y + 1, & \text{if } x - y < 0 \end{cases}$$

If  $U_1, U_2, U_3$  are independent uniform random variables over  $[0, 1)$ , then so are  $U_1 \oplus U_2, U_2 \oplus U_3$  and  $U_3 \oplus U_1$  (see Lemma 4.5 in Beer et al. (2011)). Therefore,  $\mathbf{G}(3, p_e)$

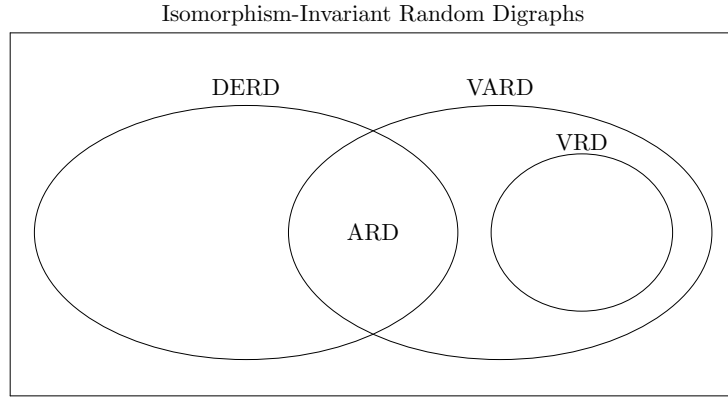


Figure 4.3: Venn diagram of DERDs and VARDs for  $n \geq 4$ . The results in this paper indicate that all the five regions in the figure are nonempty. In addition, the intersection of the classes DERDs and VARDs is the family of ARDs, i.e.,  $\text{DERD} \cap \text{VARD} = \text{ARD}$ .

can be represented as a vertex random graph  $\mathbf{G}(3, [0, 1], \nu, f)$  where  $\nu$  is the uniform distribution on  $[0, 1]$  and  $f(x, y) = \mathbf{1}_{\{x \oplus y \leq p_e\}}$  (Beer et al. (2011)).

Let  $g(x, y) = \mathbf{1}_{\{x \oplus y \leq 1/2\}} + (2p_d - 1)\mathbf{1}_{\{x \oplus y > 1/2\}}$  for every  $0 \leq x, y < 1$ . We claim that  $\mathbf{D}(3, p_e, p_d)$  is a VARD,  $\mathbf{D}(3, \Omega, \mu, \phi)$ , where  $\Omega = [0, 1] \times [0, 1]$ ,  $\mu$  is product of two uniform distributions on  $[0, 1]$  and  $\phi((u_1, u'_1), (u_2, u'_2)) = f(u_1, u_2)g(u'_1, u'_2)$ . First note that

$$g(x, y) + g(y, x) = 2p_d \quad \text{and} \quad g(x, y)g(y, x) = 2p_d - 1, \quad (4.28)$$

for every  $0 \leq x, y < 1$ . As  $f$  is a symmetric indicator function, the equations in (4.28) imply

$$\phi((u_1, u'_1), (u_2, u'_2))\phi((u_2, u'_2), (u_1, u'_1)) = f(u_1, u_2)(2p_d - 1), \quad (4.29)$$

$$\phi((u_1, u'_1), (u_2, u'_2))(1 - \phi((u_2, u'_2), (u_1, u'_1))) = f(u_1, u_2)(g(u'_1, u'_2) - (2p_d - 1)), \quad (4.30)$$

$$(1 - \phi((u_1, u'_1), (u_2, u'_2)))(1 - \phi((u_2, u'_2), (u_1, u'_1))) = 1 - f(u_1, u_2). \quad (4.31)$$

We next focus on the function  $g$ . It is easy to see that

$$\int_0^1 g(x, y) dy = \int_0^1 g(x, y) dx = \frac{1}{2} + \frac{1}{2}(2p_d - 1) = p_d, \quad (4.32)$$

for every  $0 \leq x, y < 1$ .

Consider the circle obtained by identifying the end points of the interval  $[0, 1]$  such that  $1/4$  is on the arc that starts from 0 and ends at  $1/2$  along the clockwise direction. Then,  $x \oplus y$  is equal to the length of the arc of this circle which starts from  $x$  and ends at  $y$

along the counterclockwise direction. Notice that  $g(x, y)g(y, z)$  and  $g(x, y)g(y, z)g(z, x)$  depends on the ordering of  $x, y, z$  along counterclockwise direction and whether the points  $x, y, z$  form an acute or obtuse triangle.

If  $x, y, z$  form an acute triangle, there are basically two cases for the ordering,  $x, y, z$  or  $x, z, y$ . In the first case,  $g(x, y)g(y, z) = (2p_d - 1)^2$ , and in the latter case  $g(x, y)g(y, z) = 1^2 = 1$ .

If  $x, y, z$  form an obtuse triangle, all six permutations of  $x, y, z$  ( $x, y, z$ ;  $x, z, y$ ;  $y, x, z$ ;  $y, z, x$ ;  $z, x, y$ ;  $z, y, x$ ) are possible with the point at the middle corresponding to the obtuse angle. Then, we have  $g(x, y)g(y, z) = (2p_d - 1)^2, (2p_d - 1), (2p_d - 1), (2p_d - 1), (2p_d - 1), 1$ , respectively. Moreover, it easy to show that three uniformly at random points on the circle form an acute triangle with probability  $1/4$ . Therefore, we obtain

$$\begin{aligned} \int_{[0,1]^3} g(x, y)g(y, z) \, dx dy dz &= \frac{1}{4} \cdot \frac{1}{2} ((2p_d - 1)^2 + 1) + \frac{3}{4} \cdot \frac{1}{6} ((2p_d - 1)^2 + 4(2p_d - 1) + 1) \\ &= p_d^2. \end{aligned} \quad (4.33)$$

Similarly, we have

$$\begin{aligned} \int_{[0,1]^3} g(x, y)g(y, z)g(z, x) \, dx dy dz &= \frac{1}{8} ((2p_d - 1)^3 + 1) + \frac{1}{8} (3(2p_d - 1)^2 + 3(2p_d - 1)) \\ &= p_d^3. \end{aligned} \quad (4.34)$$

By using the results in (4.28)-(4.34), one can easily verify that

$$\int P_{\mathbf{x}}(D) d(\mu_{\mathbf{x}}) = p_e^{n_e} (1 - p_e)^{3-n_e} (1 - p_d)^{n_{as}} (2p_d - 1)^{n_s},$$

for every  $D \in \mathcal{D}_3$ , and hence the desired result follows. Furthermore, the same setting works for  $n = 2$  as well.  $\square$

*Remark 4.5.5. Is every DERD with  $n = 3$  a VRD?* Note that the function  $\phi$  constructed in the proof of Theorem 4.5.4 is binary (only takes the values 0 or 1) if and only if  $p_d = 1/2$ . Hence, by Theorem 4.5.4 we see that any  $\mathbf{D}(3, p_e, 1/2)$  has a VRD representation. But, by Proposition 4.5.1,  $\mathbf{D}(3, p_e, 1/2)$  is an ARD only if  $p_e = 0$  which gives a degenerate random digraph, and hence  $\mathbf{D}(3, p_e, 1/2)$  does not yield a non-degenerate ARD. Other than the degenerate ones, is there any DERD  $\mathbf{D}(3, p_e, p_d)$  with  $p_d > 1/2$  which is also a VRD? Furthermore, is there an ARD with  $n = 3$  which has a VRD representation? For now, these questions remain to be open, but, we conjecture that any DERD with  $n = 3$  is also a VRD.

However, for  $n = 2$  the families DERDs and VRDs coincide with the all isomorphism-invariant random digraphs. Let  $D_1, D_2, D_3$  and  $D_4$  be the digraphs with vertex set  $[2]$  which only has the arc  $(1, 2)$ , only the arc  $(2, 1)$ , both of the arcs and none of the arcs,

respectively. Note that to obtain an isomorphism-invariant random digraph necessary and sufficient condition is  $P(D_1) = P(D_2)$ . Let  $\mathbf{D}$  be the random digraph with  $P(D_1) = P(D_2) = p_1$ ,  $P(D_3) = p_2$  and  $P(D_4) = 1 - 2p_1 - p_2$ . First observe that  $\mathbf{D}$  is an ARD if and only if  $\sqrt{p_2}(1 - \sqrt{p_2}) = p_1$ . Letting  $p_e = 2p_1 + p_2$  and  $p_d = (p_1 + p_2)/(2p_1 + p_2)$  gives that  $\mathbf{D}$  is a DERD  $\mathbf{D}(2, p_e, p_d)$ . With the same  $p_e$  and  $p_d$ , let  $\phi((u_1, u'_1), (u_2, u'_2)) = \mathbf{1}_{\{u_1 \oplus u'_1 \leq p_e\}} \mathbf{1}_{\{u_2 \ominus u'_2 \leq p_d\}}$ . Then, it is easy to see that  $\mathbf{D}$  is also VRD  $\mathbf{D}(2, \Omega, \mu, \phi)$ , where  $\Omega = [0, 1) \times [0, 1)$ ,  $\mu$  is the uniform distribution on  $[0, 1)^2$ . Therefore, when  $n = 2$ , any isomorphism-invariant random digraph is both a DERD and a VRD (hence also a VARD).

*Remark 4.5.6. Positive Dependence:* Recall that by the inequality in (4.20), for any VARD  $\mathbf{D}$  we have the positive dependence

$$P(\{(1, 2), (1, 3)\} \subset A(\mathbf{D})) \geq P((1, 2) \in A(\mathbf{D}))P((1, 3) \in A(\mathbf{D})) = P((1, 2) \in A(\mathbf{D}))^2. \quad (4.35)$$

Furthermore, the inequality in (4.35) can be generalized by Hölder's inequality as follows

$$P(\{(1, 2), \dots, (1, m)\} \subset A(\mathbf{D})) \geq \prod_{i=2}^m P((1, i) \in A(\mathbf{D})) = P((1, 2) \in A(\mathbf{D}))^{m-1} \quad (4.36)$$

for every VARD  $\mathbf{D}$  and  $2 \leq m \leq n$ . Similarly, we have the same inequality in (4.36) for any DVERD as well and note that equality holds for every DERD. However, there are random digraphs other than DERDs satisfying equality in (4.36) for each  $m$ . For example, consider the VRD,  $\mathbf{D}(n, [0, 1), \mu, \phi)$ , where  $\mu$  is the uniform distribution over  $[0, 1)$  and  $\phi(x, y) = \mathbf{1}_{\{x \oplus y \geq 3/8\}} \mathbf{1}_{\{y \ominus x \geq 3/8\}}$ . Clearly, in this case, we have  $P(\{(1, 2), \dots, (1, m)\} \subset A(\mathbf{D})) = (1/4)^{m-1}$  and  $P((1, i) \in A(\mathbf{D})) = 1/4$  for each  $i$ . Also, it is easy to verify that  $\mathbf{D}(n, [0, 1), \mu, \phi)$  has no DERD representation since  $P(\{(1, 2), (1, 3), (2, 3)\} \subset A(\mathbf{D})) = 0$ . In the same manner, one can easily obtain similar results for random graphs. In other words, for any VERG,  $\mathbf{G}(n, \Omega, \mu, \phi)$ , and  $2 \leq m \leq n$ , we have

$$P(\{\{1, 2\}, \{1, 3\}, \dots, \{1, m\}\} \subset E(\mathbf{G})) \geq P(\{1, 2\} \in E(\mathbf{G}))^{m-1}, \quad (4.37)$$

and equality holds for every ERG. The underlying random graph of the VRD,  $\mathbf{D}(n, [0, 1), \mu, \phi)$ , described above is an example for random graphs with no ERG representation which attains equality in (4.37) for every  $m$ .

#### 4.6 Where Do Random $k$ NN Digraphs Reside?

In this section, we determine the class relationship for random  $k$ NN digraphs. Let  $n \geq 3, k \geq 1$  and  $d \geq 1$  be integers with  $k < n - 1$ . Let  $\mu$  be a probability distribution

over  $\mathbb{R}^d$  with density function  $f$  that is assumed to be continuous almost everywhere with respect to Lebesgue measure. Let  $|\cdot|$  denote a fixed norm on  $\mathbb{R}^d$  and  $X = (X_1, \dots, X_n)$  be i.i.d. vectors in  $\mathbb{R}^d$  drawn from  $\mu$ . For given  $\mathbf{x} = (x_1, \dots, x_n)$ , let  $D_{nn}^{[k]}(\mathbf{x})$  denote the  $k$ NN digraph of  $\mathbf{x}$  with respect to the given norm  $|\cdot|$ .

**Definition 4.6.1.** The *random nearest neighbor digraph* (RNND) is the random digraph  $\mathbf{D}(n, k, d, \mu, |\cdot|)$  with

$$P(D) = \int \mathbf{1}_{\{D_{nn}^{[k]}(\mathbf{x})=D\}} d(\mu\mathbf{x}),$$

for every  $D \in \mathcal{D}_n$ .

Notice that we picked  $k$  to be less than  $n - 1$ , because otherwise, we obtain a degenerate random digraph.

**Proposition 4.6.2.** *Every RNND is isomorphism-invariant.*

*Proof.* Let  $\mathbf{D}(n, k, d, \mu, |\cdot|)$  be a RNND and  $D, D' \in \mathcal{D}_n$  be isomorphic digraphs. Then there exists a permutation  $\sigma$  on  $[n]$  such that

$$(i, j) \in A(D) \Leftrightarrow (\sigma(i), \sigma(j)) \in A(D').$$

Let  $\sigma^{-1}$  be the inverse of  $\sigma$  and  $\mathbf{y} = (y_1, \dots, y_n)$  such that  $y_i = x_{\sigma^{-1}(i)}$  for all  $1 \leq i \leq n$ , i.e.,  $y_{\sigma(i)} = x_i$ . Then it is easy to see that

$$D_{nn}^{[k]}(\mathbf{x}) = D \Leftrightarrow D_{nn}^{[k]}(\mathbf{y}) = D'.$$

The rest of the proof is similar to that of Proposition 4.2.8. □

As in VRDs, in the construction of a RNND, once  $\mathbf{x}$  is fixed, then the arcs are uniquely determined. However, in a VRD, by definition, inserting the arc  $(i, j)$  only depends on  $x_i$  and  $x_j$  whereas in a RNND it depends on all the data points. The following proposition implies that a RNND is not a VRD.

**Proposition 4.6.3.** *A RNND is neither a VARD nor a DRD.*

*Proof.* We first show that no RNND has a VARD representation. Recall that by the inequality in (4.35), for any VARD with  $n \geq 3$  we have

$$P(\{(1, 2), (1, 3)\} \subset A(\mathbf{D})) \geq (P((1, 2) \in A(\mathbf{D})))^2. \quad (4.38)$$

On the other hand, in a RNND, we have

$$\begin{aligned} P(\{(1, 2), (1, 3)\} \subset A(\mathbf{D})) &= \frac{k(k-1)}{(n-1)(n-2)} \\ &< \left(\frac{k}{n-1}\right)^2 = (P((1, 2) \in A(\mathbf{D})))^2 \end{aligned} \quad (4.39)$$

by symmetry, and hence the result follows by (4.38) and (4.39).

We show that there is no RNND which is also a DRD by contradiction. Suppose that a RNND,  $\mathbf{D} = (\mathcal{D}_n, P)$ , is a DRD generated by the random graph  $\mathbf{G} = (\mathcal{G}_n, P_{\mathbf{G}})$  and with direction probability  $p_d$ . Let  $G$  be a graph in  $\mathcal{G}_n$  with  $P_{\mathbf{G}}(G) > 0$ , and  $D$  be a digraph in  $\mathcal{D}_n$  such that  $U(D) = G$ ,  $n_s(D) = 0$  and containing a vertex which is the tail of no arc. In other words,  $D$  is a digraph whose underlying graph is  $G$ , containing no symmetric arcs and there exists a vertex  $v$  in  $V(D)$  such that  $v$  is the head of every arc incident to  $v$ . Then, as  $\mathbf{D}$  is a DRD, we have

$$P(D) = P_{\mathbf{G}}(G)(1 - p_d)^{n_{as}} > 0, \quad (4.40)$$

since  $P_{\mathbf{G}}(G) > 0$  and  $p_d < 1$ . On the other hand, for any given  $\mathbf{x} = (x_1, \dots, x_n)$ , every vertex is the tail of exactly  $k$  arcs in  $D_{nn}^{[k]}(\mathbf{x})$ . Therefore, we obtain  $P(D) = 0$  which contradicts with (4.40).  $\square$

*Remark 4.6.4.* For any set of points in  $\mathbb{R}^d$ , by space packing properties of  $\mathbb{R}^d$ , the number of points sharing a common  $k$ NN is bounded above by a constant which is independent of the number of points in the set. That is, there exists a number  $c$  which only depends on  $d, k$  and the norm  $|\cdot|$  such that in any  $k$ NN digraph, a vertex is the head of at most  $c$  arcs. Therefore, a vertex of the underlying graph of a  $k$ NN digraph is incident to at most  $c + k$  edges. Hence, if  $\mathbf{G}$  is the underlying random graph of a RNND with  $n \geq c + k + 2$ , then we have  $P(\{\{1, 2\}, \{1, 3\}, \dots, \{1, n\}\} \subset E(\mathbf{G})) = 0$  which implies that  $\mathbf{G}$  is not a VERG by the inequality in (4.37).

*Remark 4.6.5.* One can also generate NN type random digraphs other than RNNDs. For instance, in the construction of a RNND insert the arc  $(i, j)$  if and only if  $x_j$  is the  $s$ -th NN of  $x_i$  for some  $s \in S$  where  $S$  is a fixed finite nonempty set of positive integers. In other words, once the i.i.d. sample is drawn, construct the  $S$ -NN digraph of the sample. Thus, we can generalize RNNDs to  $\mathbf{D}(n, S, d, \mu, |\cdot|)$ . Then the results for RNNDs in this section are also valid for any  $\mathbf{D}(n, S, d, \mu, |\cdot|)$ , i.e., every  $\mathbf{D}(n, S, d, \mu, |\cdot|)$  is isomorphism-invariant, has no VARD or DRD representation, and for large  $n$ , has an underlying random graph which is not a VERG.

Note that Proposition 4.6.3 implies that the regions  $\text{VARD}^c$  in Figure 4.1,  $\text{DRD}^c$  in Figure 4.2 and  $(\text{DERD} \cup \text{VARD})^c$  in Figure 4.3 are nonempty.

For  $n = 3$ , the only possible value of  $k$  is 1. In this case, the pair with the minimum distance are NNs of each other and the NN of the remaining point is one of the points in this pair. Thus, by symmetry we have  $P(A(\mathbf{D}) = \{(i, j), (j, i), (k, i)\}) = 1/6$  for every pairwise distinct  $i, j, k \in \{1, 2, 3\}$ , and therefore any RNND  $\mathbf{D}(3, 1, d, \mu, |\cdot|)$  is a uniform distribution over six digraphs independent of  $d, \mu$  and  $|\cdot|$ . Also, note that the underlying random graph of  $\mathbf{D}(3, 1, d, \mu, |\cdot|)$  is always  $\mathbf{G}(3, 2)$ .

Observe that any RNND and  $\mathbf{D}(n, nk)$  have the same number of arcs. However, these

two random digraphs are different. Because, for the event  $E = \{(1, 2), \dots, (1, k+2)\} \subset A(\mathbf{D})$  we have  $P(E) = 0$  in a RNND since each vertex is tail of exactly  $k$  arcs, whereas  $P(E) > 0$  in  $\mathbf{D}(n, nk)$  since  $nk \geq k+1$ . Recall that  $n_e = n_a - n_s$ , and hence the number of edges in the underlying graph of a  $k$ NND is  $nk$  minus the number of symmetric arcs. It is easy to see that for  $n > 3$  there exist  $k$ NNDs with different number of symmetric arcs, and therefore the underlying random graph of a RNND with  $n > 3$  is not a  $\mathbf{G}(n, m)$ .



## Chapter 5

## CONCLUSIONS

In this thesis, we investigate probabilistic behavior of quantities based on  $k$ NN digraphs obtained from various random settings. We study the asymptotic behavior of subdigraph counts in  $k$ NN digraphs based on data from point processes. As point processes, we consider uniform binomial process (UBP) and homogeneous Poisson process (HPP) over a given region. For any realization of the point set, consider the  $k$ NN digraph of the data. The quantity we are interested in is the number of subdigraphs of the  $k$ NN digraph which are isomorphic to a given weakly connected digraph. The condition on the digraphs being fixed and weakly connected is crucial. Because, if the digraph is not fixed or weakly connected, then the strong stabilization condition fails in the proof of our main result. We provide LLN and CLT results for any linear combination of such quantities. In particular, we obtain asymptotic results for the number reflexive pairs, the number of shared  $k$ NNs and the number of vertices with a given indegree.

In the case of  $k = 1$  and the data is from one dimensional space (i.e.,  $d = 1$ ), by using interchangeability of uniform spacings, we provide the means and variances of the number of reflexive pairs,  $R^{[1]}$ , the number of shared  $k$ NN pairs,  $Q^{[1]}$ , and the number of vertices with a given indegree,  $Q_j^{[1]}$ . Moreover, we present a recursive relation for the probability mass function of  $R^{[1]}$ , which can provide the exact distribution of  $R^{[1]}(\mathcal{U}_n)$  (by computation for finite  $n$ ). However, the methods used for computing the variances under this setting can not be applied for higher dimensions since we lose the ordering structure, or for  $k > 1$ , since  $k$ NN relations is not determined by the ordering of the spacings when  $k > 1$  and  $d = 1$ . Also, even though simulation results strongly suggest that variances of  $R^{[1]}$  and  $Q^{[1]}$  under a binomial process with an a.e. continuous density are asymptotically  $2n/45$  and  $18n/240$ , respectively, our approach does not work to prove it since we lose the interchangeability of the spacings. Recall that the results we obtain for  $R^{[1]}(\mathcal{U}_n)$  (the mean, the variance and the recurrence relation) are in agreement with the ones in Enns et al. (1999). However, our derivation of the results are different from theirs and their method is not applicable for  $Q^{[1]}(\mathcal{U}_n)$ .

Results in Schilling (1986b) and Henze (1987) imply convergence of means results for the random variables we consider ( $R^{[k]}, Q^{[k]}$  and  $Q_j^{[k]}$ ) not only for uniform binomial process but also for any binomial point process with an almost everywhere continuous density. A potential research direction is to consider the same  $k$ NN invariants under different point processes. Our extensive Monte Carlo simulations suggest the asymptotic normality of the quantities we study whenever the underlying process has a distribution

with an a.e. continuous density.

All the asymptotic results we present have analogous versions for graphs and marked point sets as stated in Remarks 2.2.6 and 2.2.7. As an application of  $k$ NN digraphs in spatial data analysis for segregation and association, we also study  $k$ NN digraphs of data obtained by marked marked point processes, namely, marked HPP and marked UBP. We classify each arc of the  $k$ NN digraph with respect to the class membership of its tail and head, and construct the  $k$  nearest neighbor contingency table ( $k$ NNCT). Basically, the cell count denoted as  $N_{ij}$  is the number of arcs whose tail is from class  $i$  and head is from class  $j$ . For both of the underlying random processes, we derive the asymptotic values of means, variances and covariances of the cell counts and show that joint distribution of all the cell counts is asymptotically multivariate normal. We also obtain quadratic forms based on the cell counts which converge to some chi-squared distribution and determine the associated degrees of freedom as well.

Recall that in a marked HPP, if we condition on the sample size to be fixed, then we obtain a marked UBP. Moreover, if we condition on the point sets as well, we obtain another random structure that we call type 1 random labelling (RL) in which class labels of the points are independently and identically distributed for each  $n$ . Under type 1 RL, we derive the exact values of means, variances and covariances of the cell counts and prove the asymptotic joint normality of all the cell counts. Additionally, conditioned on the quantities  $r$  and  $q$  which appear in the limiting covariance values, we obtain quadratic forms (in terms of the cell counts) which have asymptotically chi-squared distributions with specified degrees of freedom. In the type 1 RL setting, if we assume that class sizes are given, then we have another type of random labelling process called type 2 RL. Hence, in type 1 RL, class sizes are random but in type 2 RL they are fixed. As in type 1 RL, we compute the exact values of means, variances and covariances of the cell counts under type 2 RL as well. But, those extra conditions on type 2 RL prevent us applying the methods used for proving asymptotic normality results which worked under marked processes and type 1 RL. Asymptotic normality of the diagonal cell counts under type 2 RL is shown by Cuzick and Edwards (1990). We prove asymptotic multivariate normality of the joint distribution of all the diagonal cell counts, however, neither our nor their approach is applicable to prove that each off-diagonal cell count under type 2 RL is asymptotically normal for data consisting of more than two classes. Yet, our simulation results strongly suggest that the distribution of the off-diagonal cell counts are also asymptotically normal. Currently, this claim remains as a conjecture and it is part of ongoing research.

Furthermore, we show that random  $k$ NN digraphs obtained by a binomial point process with an almost everywhere continuous density. are isomorphism-invariant and extend the classification of isomorphism-invariant random graphs to digraphs. Other than random  $k$ NN digraphs, we present four families, namely, arc random digraphs

(ARDs), vertex random digraphs (VRDs), vertex-arc random digraphs (VARDs) and direction random digraphs (DRDs), of isomorphism-invariant random digraphs based on where randomness resides. First three of these classes are extensions of the isomorphism-invariant random graph classes presented in Beer et al. (2011) to digraphs. The family of DRDs is obtained by randomly assigning directions to edges of isomorphism-invariant random graphs, and includes three families, direction-edge random digraphs (DERDs), direction-vertex random digraphs (DVRDs) and direction-vertex-edge random digraphs (DVERDs).

For  $n \geq 4$ , we show that there is no random digraph that is both an ARD and a VRD (which is the digraph counterpart of the result in Beer et al. (2011), that there is no non-degenerate random graph which is both an ERG and a VRG for  $n \geq 4$ ). We also show that for  $n \geq 4$ , ARDs are the only random digraphs which have both DERD and VARD representations. The method we use for the latter result is not applicable for the intersection of the families DVRDs and VARDs, since we lose the independence of the edges in a VRG. Therefore, identifying all random digraphs with both DVRD and VARD representations is a challenging problem, and remains open. We also study random  $k$ NN digraphs (RNNDs) and determine where they lie in this classification. We show that no RNND has a DRD or a VARD representation, and the underlying random graph of a RNND with large  $n$  is not a VERG. In other words, no random  $k$ NN digraph can be modeled as an DRD or a VARD.

## Appendix A

## PROOFS OF AUXILIARY RESULTS FOR THE MAIN THEOREM

## A.1 Proof of Proposition 2.2.1

Proposition 2.2.1 is a corollary of Theorem 2.1 in Penrose and Yukich (2001). We need some more notation and definitions to state their theorem.

Let  $B_r^\infty(x)$  denote the corresponding  $l_\infty$  ball, that is  $B_r^\infty(x) = [-r, r]^d + x$ . For  $r > 0$ , set  $\partial_r A = \cup_{x \in \partial A} B_r^\infty(x)$ . Let the diameter of  $A$  be  $\text{diam}(A) = \sup\{\|x - y\| : x, y \in A\}$ .

Let  $\lambda$  be a positive constant and  $(B_n)_{n \geq 1}$  be a sequence of bounded Borel subsets of  $\mathbb{R}^d$ , satisfying the following conditions:

- (1)  $m(B_n) = n/\lambda$  for all  $n$ .
- (2)  $B_n$  tends to  $\mathbb{R}^d$  as  $n \rightarrow \infty$ , i.e.,  $\cup_{n \geq 1} \cap_{m \geq n} B_m = \mathbb{R}^d$ .
- (3)  $\lim_{n \rightarrow \infty} m(\partial_r B_n)/n = 0$  for all  $r > 0$ .
- (4) There exists a constant  $\beta_1$  such that  $\text{diam}(B_n) \leq \beta_1 n^{\beta_1}$  for all  $n$ . Subject to these four conditions, the choice of  $(B_n)_{n \geq 1}$  is arbitrary.

Let  $U_{1,n}, U_{2,n}, \dots$  be i.i.d. uniform variables on  $B_n$  and  $\mathcal{U}_{m,n} = \{U_{1,n}, \dots, U_{m,n}\}$  (UBP of size  $m$  on  $B_n$ ). Let  $\mathcal{P}_{n,\lambda}$  be a HPP on  $B_n$  of intensity  $\lambda$  and  $\mathcal{P}_\lambda$  be the HPP of intensity  $\lambda$  on  $\mathbb{R}^d$ .

The functional  $H$  is called *strongly stabilizing*, if there exist a.s. finite random variables  $S$  and  $\Delta_H(\infty)$  such that with probability 1, we have

$$\Delta_H((\mathcal{P}_\lambda \cap B_S(0)) \cup V) = \Delta_H(\infty)$$

for all finite  $V \subset \mathbb{R}^d \setminus B_S(0)$ .

Let  $\mathcal{B}$  be the collection of all regions of the form  $B_n + x$  where  $x \in \mathbb{R}^d$  and  $n \geq 1$ . Given  $A \in \mathcal{B}$ , let  $\mathcal{U}_{m,A}$  be the UBP of size  $m$  on  $A$ . The functional  $H$  satisfies the *uniform bounded moments condition* on  $\mathcal{B}$ , if

$$\sup_{A \in \mathcal{B}, 0 \in A} \sup_{m \in [\lambda m(A)/2, 3\lambda m(A)/2]} \mathbf{E}(\Delta_H(\mathcal{U}_{m,A})^4) < \infty.$$

The functional  $H$  is *polynomially bounded*, if there exists a constant  $\beta$  such that for all finite sets  $V \subset \mathbb{R}^d$ ,

$$|H(V)| \leq \beta(\text{diam}(V) + \text{card}(V))^\beta.$$

**Theorem A.1.1** (Theorem 2.1 in Penrose and Yukich (2001)). *Suppose that  $H$  is*

translation-invariant, strongly stabilizing, satisfies the uniform bounded moments condition on  $\mathcal{B}$ , and is polynomially bounded. Then there exist constants  $\sigma^2$  and  $\tau^2$ , with  $0 \leq \tau^2 \leq \sigma^2$ , such that as  $n \rightarrow \infty$ ,

$$n^{-1} \mathbf{Var}(H(\mathcal{P}_{n,\lambda})) \rightarrow \sigma^2$$

and

$$n^{-1/2}(H(\mathcal{P}_{n,\lambda}) - \mathbf{E}(H(\mathcal{P}_{n,\lambda}))) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

while

$$n^{-1} \mathbf{Var}(H(\mathcal{U}_{n,n})) \rightarrow \tau^2$$

and

$$n^{-1/2}(H(\mathcal{U}_{n,n}) - \mathbf{E}(H(\mathcal{U}_{n,n}))) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau^2).$$

Also for given  $\lambda$ ,  $\sigma^2$  and  $\tau^2$  are independent of the choice of  $(B_n)_{n \geq 1}$ .

Moreover, if the distribution of  $\Delta_H(\infty)$  is non-degenerate, then  $\tau^2 > 0$  and hence also  $\sigma^2 > 0$ .

We now show that Theorem A.1.1 implies Proposition 2.2.1. By translation invariance of  $H$ , we may assume that the origin is in the interior of  $B_0$ . Let  $B_n = n^{1/d} B_0$  for each  $n \geq 1$ . Then by scale invariance, we can take  $\lambda = 1$  and  $m(B_0) = 1$ .

We check four conditions for  $(B_n)_{n \geq 1}$ . Note that  $m(B_n) = n \cdot m(B_0) = n/\lambda$  and hence, condition (1) is satisfied. Since 0 is in the interior of  $B_0$ , condition (2) holds as well. Condition (3) follows by  $m(\partial(B_0)) = 0$ . Since  $B_0$  is bounded, we also obtain the last condition and therefore,  $(B_n)_{n \geq 1}$  satisfies all the required properties.

Clearly, if  $H$  has a radius of stabilization and bounded add one cost, then it is strongly stabilizing. Moreover, since  $H$  has a bounded add one cost,  $H$  obviously satisfies the uniform bounded moments condition on  $\mathcal{B}$ . Furthermore, if  $H$  is linearly bounded, then it is polynomially bounded as well.

Finally, notice that as  $H$  is scale invariant,  $H(\mathcal{U}_{n,n})$  and  $H(\mathcal{U}_n)$  have the same distributions likewise  $H(\mathcal{P}_{n,\lambda})$  and  $H(\mathcal{P}_n)$ . Consequently, we can plug in  $\mathcal{U}_n$  and  $\mathcal{P}_n$  instead of  $\mathcal{U}_{n,n}$  and  $\mathcal{P}_{n,\lambda}$  in Theorem A.1.1, respectively, and hence, desired result follows.

## A.2 Proof of Proposition 2.2.5

Let  $(\mathcal{K}, \mathcal{F}, P)$  be a probability space. In a marked point process, we assume marks take values in  $\mathcal{K}$  and are i.i.d. with distribution  $P$ . A marked point set is of the form  $\cup_{i=1}^n \{x_i, M_{x_i}\} \subset \mathbb{R}^d \times \mathcal{K}$  where for all  $x \in \mathbb{R}^d$  the mark is denoted by  $M_x \in \mathcal{K}$ .

Let  $T_y$  be the translation operator on  $\mathbb{R}^d \times \mathcal{K}$  which maps  $(x, M_x)$  to  $(x + y, M_x)$  and  $S_c$  be the scaling operator which sends  $(x, M_x)$  to  $(cx, M_x)$  for every  $x \in \mathbb{R}^d$  and  $M_x \in \mathcal{K}$ .

Let  $h(V, v)$  be a real valued function defined for every finite marked point set  $V \subset \mathbb{R}^d$  and  $v \in V$ . We say that  $h$  is *translation invariant* if  $h(V, v) = h(T_y(V), T_y(v))$  for every  $y \in \mathbb{R}^d$ , finite marked point set  $V$  and  $v \in V$ . Similarly,  $h$  is called *scale invariant* if  $h(V, v) = h(S_y(V), S_y(v))$  for every  $c > 0$ , finite marked point set  $V$  and  $v \in V$ .

Strong stabilization of  $h$  is defined in the same way of functionals defined on unmarked point sets. Given  $p \geq 1$ ,  $h$  satisfies the *bounded  $p$ th moments condition* on  $\mathcal{B}$  at intensity  $\lambda$  if

$$\sup_{A \in \mathcal{B}, 0 \in A} \sup_{m \in [\lambda m(A)/2, 3\lambda m(A)/2]} \mathbf{E}[h(\mathcal{U}_{m,A} \cup \{0\}, 0)^p] < \infty.$$

Let  $H(V) = \sum_{v \in V} h(V, v)$ . For  $p \geq 1$ , the functional  $H$  satisfies the *bounded  $p$ th moments condition* on  $\mathcal{B}$  at intensity  $\lambda$  if

$$\sup_{A \in \mathcal{B}, 0 \in A} \sup_{m \in [\lambda m(A)/2, 3\lambda m(A)/2]} \mathbf{E}(\Delta_H(\mathcal{U}_{m,A})^p) < \infty.$$

**Theorem A.2.1** (Theorem 3.2 in Penrose and Yukich (2002)). *Assume that  $H$  is polynomially bounded with index  $\beta$  and there exists  $p > 2(1 + \beta)$  such that  $H$  satisfies the bounded  $p$ th moments condition on  $\mathcal{B}$  at intensity  $\lambda$ . Suppose also that  $h$  is strongly stabilizing with respect to  $\mathcal{P}_\lambda$ , with limit  $h_\infty$ , and  $h$  satisfies the bounded  $q$ th moments condition on  $\mathcal{B}$  at intensity  $\lambda$ , for some  $q > 1$ . Then as  $n \rightarrow \infty$ ,*

$$n^{-1}H(\mathcal{U}_{[\lambda n], n}) \xrightarrow{c.m.c.c.} \lambda \mathbf{E}(h_\infty) \text{ and } n^{-1}H(\mathcal{P}_{n, \lambda}) \xrightarrow{c.m.c.c.} \lambda \mathbf{E}(h_\infty),$$

where  $[x]$  denotes the largest integer not greater than  $x$ .

We prove that Theorem A.2.1 implies Proposition 2.2.5. Again by scale invariance and translation invariance, we make the same construction and assumptions for  $(B_n)_{n \geq 1}$  and  $\lambda$ , that is, 0 is in the interior of  $B_0$ ,  $B_n = n^{1/d}B_0$ ,  $m(B_0) = 1$  and  $\lambda = 1$ .

Since  $h$  is uniformly bounded, there exists a constant  $c$  such that  $|h(V, v)| \leq c$  for all marked point set  $V$  and  $v \in V$ . Therefore, as  $H(V) = \sum_{v \in V} h(V, v)$ , we see that  $H$  is polynomially bounded with index  $c$ . Moreover, since  $H$  has a bounded add one cost,  $H$  satisfies bounded  $p$ th moments condition for any  $p > 2(1 + c)$ . Since  $h$  is uniformly bounded,  $h$  satisfies the bounded  $q$ th moments condition for every  $q \geq 1$ . And finally, by scale invariance, we can plug in  $\mathcal{U}_n$  and  $\mathcal{P}_n$  instead of  $\mathcal{U}_{[\lambda n], n}$  and  $\mathcal{P}_{n, \lambda}$ , respectively, and hence desired result follows.

## Appendix B

DERIVATION OF MEANS AND VARIANCES OF  $R^{[1]}$  AND  $Q^{[1]}$ **B.1 Proof of Theorem 2.6.2**

By (2.32) it suffices to derive the mean and the variance of  $\sum_{i=1}^{n-1} \mathbf{1}_{A_i}$ . We first compute the mean. Clearly,  $\mathbf{E}(\mathbf{1}_{A_1}) = P(A_1) = P(\sigma(2) < \sigma(1)) = 1/2$  and similarly by symmetry  $\mathbf{E}(\mathbf{1}_{A_{n-1}}) = 1/2$ . For  $2 \leq i \leq n-2$ , we easily get  $\mathbf{E}(\mathbf{1}_{A_i}) = P(A_i) = P(\sigma(i) < \min\{\sigma(i-1), \sigma(i+1)\}) = 1/3$ . Thus, for  $n \geq 3$  we obtain

$$\mathbf{E}(R^{[1]}(\mathcal{U}_n)) = \mathbf{E}\left(\sum_{i=1}^{n-1} \mathbf{1}_{A_i}\right) = \sum_{i=1}^{n-1} \mathbf{E}(\mathbf{1}_{A_i}) = \frac{1}{2} + (n-3)\frac{1}{3} + \frac{1}{2} = \frac{n}{3}.$$

For the variance of  $R^{[1]}(\mathcal{U}_n)$ , we derive the covariances of  $\mathbf{1}_{A_i}$ 's given in the following matrix:

$$(\mathbf{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}))_{i,j=1}^{n-1} = \begin{pmatrix} \frac{1}{4} & \frac{-1}{6} & \frac{1}{24} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{-1}{6} & \frac{2}{9} & \frac{-1}{9} & \frac{1}{45} & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{24} & \frac{-1}{9} & \frac{2}{9} & \frac{-1}{9} & \frac{1}{45} & 0 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{45} & \frac{-1}{9} & \frac{2}{9} & \frac{-1}{9} & \frac{1}{45} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{45} & \frac{-1}{9} & \frac{2}{9} & \frac{-1}{9} & \frac{1}{45} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \frac{1}{45} & \frac{-1}{9} & \frac{2}{9} & \frac{-1}{9} & \frac{1}{45} & 0 \\ 0 & 0 & \cdots & & 0 & \frac{1}{45} & \frac{-1}{9} & \frac{2}{9} & \frac{-1}{9} & \frac{1}{24} \\ 0 & 0 & \cdots & & & 0 & \frac{1}{45} & \frac{-1}{9} & \frac{2}{9} & \frac{-1}{6} \\ 0 & 0 & \cdots & & & & 0 & \frac{1}{24} & \frac{-1}{6} & \frac{1}{4} \end{pmatrix}.$$

First notice that the events  $A_i$  and  $A_j$  are independent whenever  $|i-j| > 2$ , since each  $A_i$  only depends on the ordering of  $\sigma(i-1), \sigma(i)$  and  $\sigma(i+1)$ . Thus, we get  $\mathbf{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) = 0$  if  $|i-j| > 2$ . Remaining covariances on the diagonal strip  $|i-j| \leq 2$  are computed as follows. By symmetry assume  $i \leq j$ . For  $i = j$ , one can easily have

$$\mathbf{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_i}) = \mathbf{Var}(\mathbf{1}_{A_i}) = P(A_i)(1 - P(A_i)) = \begin{cases} 1/4 & i = 1 \text{ or } n-1, \\ 2/9 & 2 \leq i \leq n-2. \end{cases} \quad (\text{B.1})$$

Next, we compute the off diagonal terms on the strip  $|i - j| \leq 2$ . For  $i = 1$ , we have

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{A_1}, \mathbf{1}_{A_2}) &= P(A_1 \cap A_2) - \frac{1}{2} \cdot \frac{1}{3} = P(\sigma(1) < \sigma(2), \sigma(2) < \sigma(1), \sigma(2) < \sigma(3)) - \frac{1}{6} \\ &= 0 - \frac{1}{6} = -\frac{1}{6},\end{aligned}\tag{B.2}$$

since the event  $\{\sigma(1) < \sigma(2), \sigma(2) < \sigma(1), \sigma(2) < \sigma(3)\}$  can not occur and

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{A_1}, \mathbf{1}_{A_3}) &= P(A_1 \cap A_3) - \frac{1}{2} \cdot \frac{1}{3} = P(\sigma(1) < \sigma(2) > \sigma(3) < \sigma(4)) - \frac{1}{6} \\ &= \frac{5}{24} - \frac{1}{6} = \frac{1}{24}.\end{aligned}\tag{B.3}$$

where  $5/24$  comes from the fact that there are 5 alternating permutations of order 4. By symmetry, we also have

$$\mathbf{Cov}(\mathbf{1}_{A_{n-2}}, \mathbf{1}_{A_{n-1}}) = -1/6 \text{ and } \mathbf{Cov}(\mathbf{1}_{A_{n-3}}, \mathbf{1}_{A_{n-1}}) = 1/24.\tag{B.4}$$

When  $j = i + 1$ , for each  $2 \leq i \leq n - 3$  we have

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_{i+1}}) &= \mathbf{Cov}(\mathbf{1}_{A_2}, \mathbf{1}_{A_3}) = P(A_2 \cap A_3) - \frac{1}{3} \cdot \frac{1}{3} \\ &= P(\sigma(2) < \sigma(1), \sigma(2) < \sigma(3), \sigma(3) < \sigma(2), \sigma(3) < \sigma(4)) - \frac{1}{9} = 0 - \frac{1}{9} \\ &= -\frac{1}{9},\end{aligned}\tag{B.5}$$

since the event  $\{\sigma(2) < \sigma(1), \sigma(2) < \sigma(3), \sigma(3) < \sigma(2), \sigma(3) < \sigma(4)\}$  can not occur. And, finally, when  $j = i + 2$ , for each  $2 \leq i \leq n - 4$  we obtain

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_{i+2}}) &= \mathbf{Cov}(\mathbf{1}_{A_2}, \mathbf{1}_{A_4}) = P(A_2 \cap A_4) - \frac{1}{3} \cdot \frac{1}{3} \\ &= P(\sigma(1) > \sigma(2) < \sigma(3) > \sigma(4) < \sigma(5)) - \frac{1}{9} \\ &= \frac{16}{120} - \frac{1}{9} = \frac{1}{45}.\end{aligned}\tag{B.6}$$

where  $16/120$  comes from the fact that there are 16 alternating permutations of order 5. Therefore, for each  $n \geq 5$ , by combining the equations in (B.1)-(B.6) we obtain

$$\begin{aligned}\mathbf{Var}(R^{[1]}(\mathcal{U}_n)) &= \mathbf{Var}\left(\sum_{i=1}^{n-1} \mathbf{1}_{A_i}\right) = \sum_{i,j=1}^{n-1} \mathbf{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) \\ &= 2 \cdot \frac{1}{4} + (n-3) \frac{2}{9} + 4 \cdot \frac{-1}{6} + 4 \cdot \frac{1}{24} + 2(n-4) \frac{-1}{9} + 2(n-5) \frac{1}{45} \\ &= \frac{2n}{45}.\end{aligned}$$

### B.2 Proof of Corollary 2.6.3

For any real number  $s$ , we have

$$\begin{aligned}
 \mathbf{E}((R^{[1]}(\mathcal{P}_n))^s) &= \sum_{k=0}^{\infty} k^s P(R^{[1]}(\mathcal{P}_n) = k) = \sum_{k=0}^{\infty} k^s \sum_{m=0}^{\infty} P(\text{card}(\mathcal{P}_n) = m) P(R^{[1]}(\mathcal{U}_m) = k) \\
 &= \sum_{m=0}^{\infty} P(\text{card}(\mathcal{P}_n) = m) \sum_{k=0}^{\infty} k^s P(R^{[1]}(\mathcal{U}_m) = k) \\
 &= \sum_{m=0}^{\infty} P(\text{card}(\mathcal{P}_n) = m) \mathbf{E}((R^{[1]}(\mathcal{U}_m))^s)
 \end{aligned} \tag{B.7}$$

Note that  $\mathbf{E}(R^{[1]}(\mathcal{U}_0)) = \mathbf{E}(R^{[1]}(\mathcal{U}_1)) = 0$ ,  $\mathbf{E}(R^{[1]}(\mathcal{U}_2)) = 1$  and  $\mathbf{E}(R^{[1]}(\mathcal{U}_m)) = m/3$  for every  $m \geq 3$ . Then letting  $s = 1$  in (B.7) yields

$$\begin{aligned}
 \mathbf{E}(R^{[1]}(\mathcal{P}_n)) &= \sum_{m=0}^{\infty} \frac{n}{3} P(\text{card}(\mathcal{P}_n) = m) - \frac{1}{3} P(\text{card}(\mathcal{P}_n) = 1) + \frac{1}{3} P(\text{card}(\mathcal{P}_n) = 2) \\
 &= \frac{\mathbf{E}(\text{card}(\mathcal{P}_n))}{3} + \frac{e^{-n}n(n-2)}{6} \\
 &= \frac{n}{3} + \frac{e^{-n}n(n-2)}{6},
 \end{aligned} \tag{B.8}$$

since  $\mathbf{E}(\text{card}(\mathcal{P}_n)) = n$ . It is easy to verify  $\mathbf{E}((R^{[1]}(\mathcal{U}_0))^2) = \mathbf{E}((R^{[1]}(\mathcal{U}_1))^2) = 0$ ,  $\mathbf{E}((R^{[1]}(\mathcal{U}_2))^2) = \mathbf{E}((R^{[1]}(\mathcal{U}_3))^2) = 1$  and  $\mathbf{E}((R^{[1]}(\mathcal{U}_4))^2) = 2$ . As  $\mathbf{E}((R^{[1]}(\mathcal{U}_m))^2) = (5m^2 + 2m)/45$  for every  $m \geq 5$ , letting  $s = 2$  in (B.7) gives

$$\begin{aligned}
 \mathbf{E}((R^{[1]}(\mathcal{P}_n))^2) &= \sum_{m=0}^{\infty} \frac{5m^2 + 2m}{45} P(\text{card}(\mathcal{P}_n) = m) - \frac{7}{45} P(\mathcal{P}_n = 1) \\
 &\quad + \frac{21}{45} P(\text{card}(\mathcal{P}_n) = 2) - \frac{6}{45} P(\text{card}(\mathcal{P}_n) = 3) + \frac{2}{45} P(\text{card}(\mathcal{P}_n) = 4) \\
 &= \frac{5\mathbf{E}((\text{card}(\mathcal{P}_n))^2) + 2\mathbf{E}(\text{card}(\mathcal{P}_n))}{45} + \frac{e^{-n}n(n^3 - 12n^2 + 132n - 84)}{540} \\
 &= \frac{5n^2 + 7n}{45} + \frac{e^{-n}n(n^3 - 12n^2 + 132n - 84)}{540},
 \end{aligned} \tag{B.9}$$

since  $\mathbf{E}(\text{card}(\mathcal{P}_n)) = n$  and  $\mathbf{E}((\text{card}(\mathcal{P}_n))^2) = n^2 + n$ . Thus, by (B.8) and (B.9) we get

$$\mathbf{Var}(R^{[1]}(\mathcal{P}_n)) = \frac{7n}{45} + \frac{e^{-n}n(n^3 - 72n^2 + 252n - 84)}{540} - \frac{e^{-2n}n^2(n-2)^2}{36}.$$

### B.3 Proof of Theorem 2.6.4

Again by (2.32), we compute the mean and the variance of  $\sum_{i=1}^{n-2} \mathbf{1}_{B_i}$ . Clearly,

$$\mathbf{E}(\mathbf{1}_{B_i}) = P(B_i) = \begin{cases} 1/2 & i = 1 \text{ or } n-2, \\ 1/4 & 2 \leq i \leq n-3, \end{cases}$$

and hence, for every  $n \geq 4$ , we get

$$\mathbf{E}(Q^{[1]}(\mathcal{U}_n)) = \mathbf{E}\left(\sum_{i=1}^{n-2} \mathbf{1}_{B_i}\right) = \sum_{i=1}^{n-2} \mathbf{E}(\mathbf{1}_{B_i}) = \frac{1}{2} + (n-4)\frac{1}{4} + \frac{1}{2} = \frac{n}{4}.$$

For the variance of  $Q^{[1]}(\mathcal{U}_n)$ , we compute the covariances of  $\mathbf{1}_{B_i}$ 's in the following matrix  $(\mathbf{Cov}(\mathbf{1}_{B_i}, \mathbf{1}_{B_j}))_{i,j=1}^{n-2}$  which is equal to

$$\begin{pmatrix} \frac{1}{4} & 0 & -\frac{1}{8} & \frac{1}{24} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{48} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -\frac{1}{8} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{48} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{24} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{48} & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{48} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{48} & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{48} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{48} & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \frac{1}{48} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{48} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \frac{1}{48} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{48} & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{48} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{24} \\ 0 & 0 & \cdots & 0 & \frac{1}{48} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & -\frac{1}{16} & \frac{1}{8} \\ 0 & 0 & \cdots & 0 & \frac{1}{48} & -\frac{1}{16} & -\frac{1}{80} & \frac{3}{16} & -\frac{1}{80} & \frac{1}{16} & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{24} & -\frac{1}{8} & 0 & \frac{1}{4} \end{pmatrix}.$$

Note that the events  $B_i$  and  $B_j$  are independent whenever  $|i - j| > 3$ , since each  $B_i$  only depends on the ordering of  $\sigma(i-1), \sigma(i), \sigma(i+1)$  and  $\sigma(i+2)$ . Therefore, we have  $\mathbf{Cov}(\mathbf{1}_{B_i}, \mathbf{1}_{B_j}) = 0$  if  $|i - j| > 3$ . Remaining covariances (i.e., the entries in the diagonal strip  $|i - j| \leq 3$ ) are computed as follows. By symmetry suppose  $i \leq j$ . When  $i = j$ , one can easily obtain the main diagonal terms

$$\mathbf{Cov}(\mathbf{1}_{B_i}, \mathbf{1}_{B_i}) = \mathbf{Var}(\mathbf{1}_{B_i}) = P(B_i)(1 - P(B_i)) = \begin{cases} 1/4 & i = 1 \text{ or } n-2, \\ 3/16 & 2 \leq i \leq n-3. \end{cases} \quad (\text{B.10})$$

We next compute the off diagonal terms on the strip. For  $i = 1$ , we have

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{B_1}, \mathbf{1}_{B_2}) &= P(B_1 \cap B_2) - \frac{1}{2} \cdot \frac{1}{4} = P(\sigma(2) < \sigma(3), \sigma(2) < \sigma(1), \sigma(3) < \sigma(4)) - \frac{1}{8} \\ &= P(\sigma(2) < \min\{\sigma(1), \sigma(3), \sigma(4)\}, \sigma(3) < \sigma(4)) - \frac{1}{8} = \frac{1}{4} \cdot \frac{1}{2} - \frac{1}{8} \\ &= 0,\end{aligned}\tag{B.11}$$

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{B_1}, \mathbf{1}_{B_3}) &= P(B_1 \cap B_3) - \frac{1}{2} \cdot \frac{1}{4} = P(\sigma(2) < \sigma(3), \sigma(3) < \sigma(2), \sigma(4) < \sigma(5)) - \frac{1}{8} \\ &= 0 - \frac{1}{8} = -\frac{1}{8},\end{aligned}\tag{B.12}$$

since the event  $\{\sigma(2) < \sigma(3), \sigma(3) < \sigma(2), \sigma(4) < \sigma(5)\}$  can not occur and

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{B_1}, \mathbf{1}_{B_4}) &= P(B_1 \cap B_4) - \frac{1}{2} \cdot \frac{1}{4} = P(\sigma(2) < \sigma(3), \sigma(4) < \sigma(3), \sigma(5) < \sigma(6)) - \frac{1}{8} \\ &= P(\max\{\sigma(2), \sigma(4)\} < \sigma(3), \sigma(5) < \sigma(6)) - \frac{1}{8} = \frac{1}{3} \cdot \frac{1}{2} - \frac{1}{8} \\ &= \frac{1}{24}.\end{aligned}\tag{B.13}$$

By symmetry, we get

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{B_{n-3}}, \mathbf{1}_{B_{n-2}}) &= 0, \mathbf{Cov}(\mathbf{1}_{B_{n-4}}, \mathbf{1}_{B_{n-2}}) = -1/8 \text{ and } \mathbf{Cov}(\mathbf{1}_{B_{n-5}}, \mathbf{1}_{B_{n-2}}) \\ &= 1/24.\end{aligned}\tag{B.14}$$

When  $j = i + 1$  and  $2 \leq i \leq n - 4$ , we obtain

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{B_i}, \mathbf{1}_{B_{i+1}}) &= \mathbf{Cov}(\mathbf{1}_{B_2}, \mathbf{1}_{B_3}) = P(B_2 \cap B_3) - \frac{1}{4} \cdot \frac{1}{4} \\ &= P(\sigma(2) < \sigma(1), \sigma(3) < \sigma(4), \sigma(3) < \sigma(2), \sigma(4) < \sigma(5)) - \frac{1}{16} \\ &= P(\sigma(3) < \min\{\sigma(1), \sigma(2), \sigma(4), \sigma(5)\}, \sigma(2) < \sigma(1), \sigma(4) < \sigma(5)) - \frac{1}{16} \\ &= \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} - \frac{1}{16} = \frac{-1}{80}.\end{aligned}\tag{B.15}$$

Similarly, if  $j = i + 2$  and  $2 \leq i \leq n - 5$  we have

$$\begin{aligned}\mathbf{Cov}(\mathbf{1}_{B_i}, \mathbf{1}_{B_{i+2}}) &= \mathbf{Cov}(\mathbf{1}_{B_2}, \mathbf{1}_{B_4}) = P(B_2 \cap B_4) - \frac{1}{4} \cdot \frac{1}{4} \\ &= P(\sigma(2) < \sigma(1), \sigma(3) < \sigma(4), \sigma(4) < \sigma(3), \sigma(5) < \sigma(6)) - \frac{1}{16} \\ &= 0 - \frac{1}{16} = -\frac{1}{16},\end{aligned}\tag{B.16}$$

since the event  $\{\sigma(2) < \sigma(1), \sigma(3) < \sigma(4), \sigma(4) < \sigma(3), \sigma(5) < \sigma(6)\}$  can not occur. Finally, for the case  $j = i + 3$  and  $2 \leq i \leq n - 6$ , we get

$$\begin{aligned}
\mathbf{Cov}(\mathbf{1}_{B_i}, \mathbf{1}_{B_{i+3}}) &= \mathbf{Cov}(\mathbf{1}_{B_2}, \mathbf{1}_{B_5}) = P(B_2 \cap B_5) - \frac{1}{4} \cdot \frac{1}{4} \\
&= P(\sigma(2) < \sigma(1), \sigma(3) < \sigma(4), \sigma(5) < \sigma(4), \sigma(6) < \sigma(7)) - \frac{1}{16} \\
&= P(\max\{\sigma(3), \sigma(5)\} < \sigma(4), \sigma(2) < \sigma(1), \sigma(6) < \sigma(7)) - \frac{1}{16} \\
&= \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} - \frac{1}{16} = \frac{1}{48},
\end{aligned} \tag{B.17}$$

and therefore, for every  $n \geq 7$ , by combining the equations in (B.10)-(B.17) we obtain

$$\begin{aligned}
\mathbf{Var}(Q^{[1]}(\mathcal{U}_n)) &= \mathbf{Var}\left(\sum_{i=1}^{n-2} \mathbf{1}_{B_i}\right) = \sum_{i,j=1}^{n-2} \mathbf{Cov}(\mathbf{1}_{B_i}, \mathbf{1}_{B_j}) \\
&= 2 \cdot \frac{1}{4} + (n-4) \frac{3}{16} + 4 \cdot \frac{-1}{8} + 4 \cdot \frac{1}{24} + 2(n-5) \frac{-1}{80} + 2(n-6) \frac{-1}{16} \\
&\quad + 2(n-7) \frac{1}{48} \\
&= \frac{19n}{240}.
\end{aligned}$$

Remaining part of the proof, derivation of mean and variance of  $Q^{[1]}(\mathcal{P}_n)$ , can be easily completed by the same ideas in the proof of Corollary 2.6.3 (see, Appendix B.2).

## Appendix C

### DERIVATION OF MEANS, VARIANCES AND COVARIANCES OF THE CELL COUNTS IN $kNNCT$

#### C.1 Under $(\mathcal{P}_n, \mathcal{M})$

We compute the means, variances and covariances of  $N_{\alpha\beta}$ 's by writing them as a linear combination of indicator random variables. Let  $N$  be the number of points in  $\mathcal{X}$ . Notice that  $N$  has a Poisson distribution with mean  $n$ . Let  $N_{\alpha\beta,s}$  be the conditional random variable of  $N_{\alpha\beta}$  given  $N = s$ . So, we clearly have

$$N_{\alpha\beta} = \sum_{s=0}^{\infty} \mathbf{1}_{\{N=s\}} N_{\alpha\beta,s} \quad (\text{C.1})$$

and

$$\mathbf{E}(N_{\alpha\beta}) = \sum_{s=0}^{\infty} P(N = s) \mathbf{E}(N_{\alpha\beta,s}).$$

Note that  $N_{\alpha\beta,s}$  is the number of arcs whose tail is marked  $\alpha$  and head is marked  $\beta$  in the  $kNN$  digraph of the sample obtained by the marked binomial point process  $(\mathcal{U}_s, \mathcal{M})$  where  $\mathcal{U}_s$  is the i.i.d. sequence of size  $s$  uniformly distributed over the unit square.

We first compute the expectation of  $N_{\alpha\beta,s}$ . For any given  $s$ , let

$$y_{ij,s} = \begin{cases} 1 & \text{if } u_j \text{ is a } kNN \text{ of } u_i, \\ 0 & \text{otherwise ;} \end{cases}$$

and for every  $1 \leq \alpha, \beta \leq m$ , let

$$z_{ij,s}^{\alpha\beta} = \begin{cases} 1 & \text{if mark of } u_i \text{ is } \alpha \text{ and mark of } u_j \text{ is } \beta, \\ 0 & \text{otherwise.} \end{cases}$$

Then, it is easy to see that

$$N_{\alpha\beta,s} = \sum_{1 \leq i \neq j \leq s} y_{ij,s} z_{ij,s}^{\alpha\beta},$$

and therefore, we obtain

$$\mathbf{E}(N_{\alpha\beta,s}) = \sum_{1 \leq i \neq j \leq s} \mathbf{E}(y_{ij,s}) \mathbf{E}(z_{ij,s}^{\alpha\beta}),$$

since  $y_{ij,s}$  and  $z_{ij,s}^{\alpha\beta}$  are independent. Note that

$$\mathbf{E}(y_{ij,s}) = \begin{cases} \frac{k}{s-1} & \text{if } s \geq k+1, \\ 1 & \text{if } 0 \leq s \leq k; \end{cases}$$

and

$$\mathbf{E}(z_{ij,s}^{\alpha\beta}) = \begin{cases} p_\alpha p_\beta & \text{if } s \geq 2, \\ 0 & \text{if } 0 \leq s \leq 1. \end{cases}$$

Therefore, for every  $s \geq k+1$  we obtain

$$\mathbf{E}(N_{\alpha\beta,s}) = s(s-1) \frac{k}{s-1} p_\alpha p_\beta = skp_\alpha p_\beta$$

and hence

$$\begin{aligned} \mathbf{E}(N_{\alpha\beta}) &= \sum_{s=0}^{\infty} P(N=s) skp_\alpha p_\beta + \sum_{s=0}^k P(N=s) (\mathbf{E}(N_{\alpha\beta,s}) - skp_\alpha p_\beta) \\ &= \mathbf{E}(N) kp_\alpha p_\beta + \epsilon_n \\ &= nk p_\alpha p_\beta + \epsilon_n, \end{aligned} \tag{C.2}$$

where

$$\epsilon_n = \sum_{s=0}^k P(N=s) (\mathbf{E}(N_{\alpha\beta,s}) - skp_\alpha p_\beta). \tag{C.3}$$

Moreover, it is easy to see that  $\epsilon_n$  is  $\mathcal{O}(n^k e^{-n})$  and thus,  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ .

To compute the variance and covariances of the cell counts, we first need to find  $\mathbf{E}(N_{\alpha\beta} N_{\theta\gamma})$ . By (C.1), we have

$$\begin{aligned} N_{\alpha\beta} N_{\theta\gamma} &= \sum_{s=0}^{\infty} \mathbf{1}_{\{N=s\}} N_{\alpha\beta,s} \sum_{s'=0}^{\infty} \mathbf{1}_{\{N=s'\}} N_{\theta\gamma,s'} \\ &= \sum_{s=0}^{\infty} \mathbf{1}_{\{N=s\}} N_{\alpha\beta,s} N_{\theta\gamma,s}, \end{aligned}$$

since  $\mathbf{1}_{\{N=s\}}\mathbf{1}_{\{N=s'\}} = \mathbf{1}_{\{N=s=s'\}}$ . Thus, we get

$$\mathbf{E}(N_{\alpha\beta}N_{\theta\gamma}) = \sum_{s=0}^{\infty} P(N=s) \mathbf{E}(N_{\alpha\beta,s}N_{\theta\gamma,s}). \quad (C.4)$$

Therefore, we are supposed to compute  $\mathbf{E}(N_{\alpha\beta,s}N_{\theta\gamma,s})$  for every  $s$ . Assume that  $s \geq \max\{4, k+1\}$ .

$$\begin{aligned} \mathbf{E}(N_{\alpha\beta,s}N_{\theta\gamma,s}) &= \mathbf{E} \left( \sum_{1 \leq i \neq j \leq s} y_{ij,s} z_{ij,s}^{\alpha\beta} \sum_{1 \leq l \neq t \leq s} y_{lt,s} z_{lt,s}^{\theta\gamma} \right) \\ &= \sum_{1 \leq i \neq j, l \neq t \leq s} \mathbf{E}(y_{ij,s} y_{lt,s}) \mathbf{E}(z_{ij,s}^{\alpha\beta} z_{lt,s}^{\theta\gamma}), \end{aligned} \quad (C.5)$$

by the linearity of the expectation and the fact that  $y_{ij,s} y_{lt,s}$  and  $z_{ij,s}^{\alpha\beta} z_{lt,s}^{\theta\gamma}$  are independent. Observe that  $\mathbf{E}(z_{ij,s}^{\alpha\beta} z_{lt,s}^{\theta\gamma})$  only depends on the concurrences between  $i, j, l, t$  for given  $\alpha, \beta, \theta, \gamma$  and hence, it is independent of  $s$  whenever  $s \geq 4$ . So, we have  $\mathbf{E}(z_{ij,s}^{\alpha\beta} z_{lt,s}^{\theta\gamma}) = \mathbf{E}(z_{ij,s'}^{\alpha\beta} z_{lt,s'}^{\theta\gamma})$  for every  $s, s' \geq 4$  and let that expectation be denoted as  $\mathbf{E}(z_{ij}^{\alpha\beta} z_{lt}^{\theta\gamma})$ . Therefore, by the linearity of expectation, the sum in (C.5) can be rewritten as sum of seven parts which are grouped with respect to the indices of  $y_{ij} y_{lt}$  as follows;

$$\begin{aligned} &\mathbf{E}(z_{12}^{\alpha\beta} z_{12}^{\theta\gamma}) \mathbf{E} \left( \sum_{i \neq j} y_{ij,s}^2 \right) + \mathbf{E}(z_{12}^{\alpha\beta} z_{21}^{\theta\gamma}) \mathbf{E} \left( \sum_{i \neq j} y_{ij,s} y_{ji,s} \right) + \mathbf{E}(z_{12}^{\alpha\beta} z_{13}^{\theta\gamma}) \mathbf{E} \left( \sum_{i \neq j \neq l} y_{ij,s} y_{il,s} \right) \\ &+ \mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) \mathbf{E} \left( \sum_{i \neq j \neq l} y_{ij,s} y_{li,s} \right) + \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) \mathbf{E} \left( \sum_{i \neq j \neq l} y_{ij,s} y_{jl,s} \right) \\ &+ \mathbf{E}(z_{12}^{\alpha\beta} z_{32}^{\theta\gamma}) \mathbf{E} \left( \sum_{i \neq j \neq l} y_{ij,s} y_{lj,s} \right) + \mathbf{E}(z_{12}^{\alpha\beta} z_{34}^{\theta\gamma}) \mathbf{E} \left( \sum_{i \neq j \neq l \neq t} y_{ij,s} y_{lt,s} \right), \end{aligned} \quad (C.6)$$

where sums are over 1 through  $s$ ,  $i \neq j \neq l$  means that  $i, j, l$  are pairwise distinct and  $i \neq j \neq l \neq t$  means that  $i, j, l, t$  are pairwise distinct. We now compute each part in (C.6).

$$\sum_{i \neq j} y_{ij,s}^2 = sk, \quad (C.7)$$

since  $y_{ij,s}^2 = y_{ij,s}$  and  $\sum_{j=1, j \neq i}^s y_{ij} = k$  for each  $i$ . Note that

$$\sum_{i \neq j} y_{ij,s} y_{ji,s} = 2R_s^{[k]}, \quad (C.8)$$

twice the number of reflexive  $kNN$ 's.

$$\sum_{i \neq j \neq l} y_{ij,s} y_{il,s} = sk(k-1), \quad (C.9)$$

since  $\sum_{1 \leq j \neq l \leq s} y_{ij,s} y_{il,s} = k(k-1)$  for each  $i$ .

$$\sum_{i \neq j \neq l} y_{ij,s} y_{li,s} = sk^2 - 2R_s^{[k]}, \quad (C.10)$$

because we choose  $l$  in  $s$  ways, a  $kNN$  of  $u_l$  (say  $u_i$ ) in  $k$  ways and a  $kNN$  of  $u_i$  (say  $u_j$ ) in  $k$  ways as well, and we subtract the number of cases in which  $j = l$ .

$$\sum_{i \neq j \neq l} y_{ij,s} y_{jl,s} = sk^2 - 2R_s^{[k]}, \quad (C.11)$$

by the same arguments in (C.10). Note that

$$\sum_{i \neq j \neq l} y_{ij,s} y_{lj,s} = 2Q_s^{[k]}, \quad (C.12)$$

twice the number of shared  $kNN$ 's.

$$\begin{aligned} \sum_{i \neq j \neq l \neq t} y_{ij,s} y_{lt,s} &= (sk)^2 - (sk + 2R_s^{[k]} + sk(k-1) + 2(sk^2 - 2R_s^{[k]}) + 2Q_s^{[k]}) \\ &= s^2k^2 - 3sk^2 + 2R_s^{[k]} - 2Q_s^{[k]}, \end{aligned} \quad (C.13)$$

since the summations in (C.7)-(C.13) add up to  $(\sum_{i \neq j} y_{ij,s})^2 = (sk)^2$ . Then, by the results in (C.6)-(C.13), we see that

$$\begin{aligned} \mathbf{E}(N_{\alpha\beta,s} N_{\theta\gamma,s}) &= \mathbf{E}(z_{12}^{\alpha\beta} z_{12}^{\theta\gamma}) sk + \mathbf{E}(z_{12}^{\alpha\beta} z_{21}^{\theta\gamma}) 2\mathbf{E}(R_s^{[k]}) + \mathbf{E}(z_{12}^{\alpha\beta} z_{13}^{\theta\gamma}) sk(k-1) \\ &\quad + \mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) (sk^2 - 2\mathbf{E}(R_s^{[k]})) + \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) (sk^2 - 2\mathbf{E}(R_s^{[k]})) \\ &\quad + \mathbf{E}(z_{12}^{\alpha\beta} z_{32}^{\theta\gamma}) 2\mathbf{E}(Q_s^{[k]}) + \mathbf{E}(z_{12}^{\alpha\beta} z_{34}^{\theta\gamma}) (s^2k^2 - 3sk^2 + 2\mathbf{E}(R_s^{[k]}) - 2\mathbf{E}(Q_s^{[k]})) \\ &= s^2k^2 \mathbf{E}(z_{12}^{\alpha\beta} z_{34}^{\theta\gamma}) \\ &\quad + sk \left( \mathbf{E}(z_{12}^{\alpha\beta} z_{12}^{\theta\gamma}) + (k-1) \mathbf{E}(z_{12}^{\alpha\beta} z_{13}^{\theta\gamma}) \right. \\ &\quad \left. + k(\mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) + \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) - 3\mathbf{E}(z_{12}^{\alpha\beta} z_{34}^{\theta\gamma})) \right) \\ &\quad + 2\mathbf{E}(R_s^{[k]}) \left( \mathbf{E}(z_{12}^{\alpha\beta} z_{34}^{\theta\gamma}) + \mathbf{E}(z_{12}^{\alpha\beta} z_{21}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) \right) \\ &\quad + 2\mathbf{E}(Q_s^{[k]}) \left( \mathbf{E}(z_{12}^{\alpha\beta} z_{32}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta} z_{34}^{\theta\gamma}) \right), \end{aligned} \quad (C.14)$$

for every  $s \geq \max\{4, k+1\}$ . Let

$$R^{[k]} = \sum_{s=0}^{\infty} \mathbf{1}_{\{N=s\}} R_s^{[k]} \text{ and } Q^{[k]} = \sum_{s=0}^{\infty} \mathbf{1}_{\{N=s\}} Q_s^{[k]}$$

be the number of reflexive pairs and shared  $k\text{NN}$ 's in  $\mathcal{X}$ , respectively. By (C.4) and (C.14), we obtain that

$$\begin{aligned} \mathbf{E}(N_{\alpha\beta}N_{\theta\gamma}) &= \mathbf{E}(N^2)k^2\mathbf{E}(z_{12}^{\alpha\beta}z_{34}^{\theta\gamma}) \\ &\quad + \mathbf{E}(N)k \left( \mathbf{E}(z_{12}^{\alpha\beta}z_{12}^{\theta\gamma}) + (k-1)\mathbf{E}(z_{12}^{\alpha\beta}z_{13}^{\theta\gamma}) \right. \\ &\quad \left. + k(\mathbf{E}(z_{12}^{\alpha\beta}z_{31}^{\theta\gamma}) + \mathbf{E}(z_{12}^{\alpha\beta}z_{23}^{\theta\gamma}) - 3\mathbf{E}(z_{12}^{\alpha\beta}z_{34}^{\theta\gamma})) \right) \\ &\quad + 2\mathbf{E}(R^{[k]}) \left( \mathbf{E}(z_{12}^{\alpha\beta}z_{34}^{\theta\gamma}) + \mathbf{E}(z_{12}^{\alpha\beta}z_{21}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta}z_{31}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta}z_{23}^{\theta\gamma}) \right) \\ &\quad + 2\mathbf{E}(Q^{[k]}) \left( \mathbf{E}(z_{12}^{\alpha\beta}z_{32}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta}z_{34}^{\theta\gamma}) \right) + \delta_n \end{aligned} \quad (\text{C.15})$$

$$\begin{aligned} &= n^2k^2p_{\alpha}p_{\beta}p_{\theta}p_{\gamma} \\ &\quad + nk \left( \mathbf{E}(z_{12}^{\alpha\beta}z_{12}^{\theta\gamma}) + (k-1)\mathbf{E}(z_{12}^{\alpha\beta}z_{13}^{\theta\gamma}) \right. \\ &\quad \left. + k(\mathbf{E}(z_{12}^{\alpha\beta}z_{31}^{\theta\gamma}) + \mathbf{E}(z_{12}^{\alpha\beta}z_{23}^{\theta\gamma}) - 2p_{\alpha}p_{\beta}p_{\theta}p_{\gamma}) \right) \\ &\quad + 2\mathbf{E}(R^{[k]}) \left( p_{\alpha}p_{\beta}p_{\theta}p_{\gamma} + \mathbf{E}(z_{12}^{\alpha\beta}z_{21}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta}z_{31}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta}z_{23}^{\theta\gamma}) \right) \\ &\quad + 2\mathbf{E}(Q^{[k]}) \left( \mathbf{E}(z_{12}^{\alpha\beta}z_{32}^{\theta\gamma}) - p_{\alpha}p_{\beta}p_{\theta}p_{\gamma} \right) + \delta_n \end{aligned} \quad (\text{C.16})$$

since  $\mathbf{E}(N^2) = n^2 + n$ ,  $\mathbf{E}(N) = n$  and  $\mathbf{E}(z_{12}^{\alpha\beta}z_{34}^{\theta\gamma}) = p_{\alpha}p_{\beta}p_{\theta}p_{\gamma}$ , where  $\delta_n$  is  $\mathcal{O}(1)$  and hence,  $\delta_n/n \rightarrow 0$  as  $n \rightarrow \infty$ . By (C.2), recall that

$$\mathbf{E}(N_{\alpha\beta})\mathbf{E}(N_{\theta\gamma}) = n^2k^2p_{\alpha}p_{\beta}p_{\theta}p_{\gamma} + n\omega_n \quad (\text{C.17})$$

where  $\omega_n \rightarrow 0$  as  $n \rightarrow \infty$ . Moreover, results in Corollary 2.4.1 and Corollary 2.5.1 imply

$$\frac{\mathbf{E}(R^{[k]})}{n} \rightarrow r = r(d, k) \text{ and } \frac{\mathbf{E}(Q^{[k]})}{n} \rightarrow q = q(d, k) \quad (\text{C.18})$$

as  $n \rightarrow \infty$ .

Since  $\mathbf{Cov}(N_{\alpha\beta}, N_{\theta\gamma}) = \mathbf{E}(N_{\alpha\beta}N_{\theta\gamma}) - \mathbf{E}(N_{\alpha\beta})\mathbf{E}(N_{\theta\gamma})$ , the results in (C.16)-(C.18)

yield  $\mathbf{Cov}(N_{\alpha\beta}, N_{\theta\gamma})/n \rightarrow \rho(\alpha\beta, \theta\gamma)$  as  $n \rightarrow \infty$ , where

$$\begin{aligned}
 \rho(\alpha\beta, \theta\gamma) &= k \left( \mathbf{E}(z_{12}^{\alpha\beta} z_{12}^{\theta\gamma}) + (k-1) \mathbf{E}(z_{12}^{\alpha\beta} z_{13}^{\theta\gamma}) + k (\mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) + \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) - 2p_{\alpha}p_{\beta}p_{\theta}p_{\gamma}) \right) \\
 &\quad + 2r \left( p_{\alpha}p_{\beta}p_{\theta}p_{\gamma} + \mathbf{E}(z_{12}^{\alpha\beta} z_{21}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) - \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) \right) \\
 &\quad + 2q \left( \mathbf{E}(z_{12}^{\alpha\beta} z_{32}^{\theta\gamma}) - p_{\alpha}p_{\beta}p_{\theta}p_{\gamma} \right) \\
 &= \mathbf{E}(z_{12}^{\alpha\beta} z_{12}^{\theta\gamma})k + 2\mathbf{E}(z_{12}^{\alpha\beta} z_{21}^{\theta\gamma})r + \mathbf{E}(z_{12}^{\alpha\beta} z_{13}^{\theta\gamma})k(k-1) \\
 &\quad + \left( \mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) + \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) \right) (k^2 - 2r) \\
 &\quad + 2\mathbf{E}(z_{12}^{\alpha\beta} z_{32}^{\theta\gamma})q - 2p_{\alpha}p_{\beta}p_{\theta}p_{\gamma}(k^2 - r + q). \tag{C.19}
 \end{aligned}$$

Now, let  $a, b, c, d$  be pairwise distinct class labels. As  $\mathbf{Var}(N_{aa}) = \mathbf{Cov}(N_{aa}, N_{aa})$ , we pick  $\alpha = \beta = \theta = \gamma = a$ . Since

$$\mathbf{E}(z_{12}^{aa} z_{12}^{aa}) = \mathbf{E}(z_{12}^{aa} z_{21}^{aa}) = p_a^2$$

and

$$\mathbf{E}(z_{12}^{aa} z_{13}^{aa}) = \mathbf{E}(z_{12}^{aa} z_{31}^{aa}) = \mathbf{E}(z_{12}^{aa} z_{23}^{aa}) = \mathbf{E}(z_{12}^{aa} z_{32}^{aa}) = p_a^3,$$

by the equation (C.19) we obtain

$$\begin{aligned}
 \rho(aa, aa) &= p_a^2(k+2r) + p_a^3(k(k-1) + 2(k^2 - 2r) + 2q) - 2p_a^4(k^2 - r + q) \\
 &= p_a^2(k+2r + p_a(3k^2 - k - 4r + 2q) - 2p_a^3(k^2 - r + q))
 \end{aligned}$$

and thus, we get (3.3a).

Similarly, since  $\mathbf{Var}(N_{ab}) = \mathbf{Cov}(N_{ab}, N_{ab})$ , we take  $\alpha = \theta = a$  and  $\beta = \gamma = b$ . Then, as

$$\mathbf{E}(z_{12}^{ab} z_{12}^{ab}) = p_a p_b, \mathbf{E}(z_{12}^{ab} z_{13}^{ab}) = p_a p_b^2, \mathbf{E}(z_{12}^{ab} z_{32}^{ab}) = p_a^2 p_b$$

and

$$\mathbf{E}(z_{12}^{ab} z_{21}^{ab}) = \mathbf{E}(z_{12}^{ab} z_{31}^{ab}) = \mathbf{E}(z_{12}^{ab} z_{23}^{ab}) = 0,$$

equation (C.19) gives

$$\begin{aligned}
 \rho(ab, ab) &= p_a p_b k + p_a p_b^2 k(k-1) + 2p_a^2 p_b q - 2p_a^2 p_b^2 (k^2 - r + q) \\
 &= p_a p_b (k + p_b k(k-1) + 2p_a q - 2p_a p_b (k^2 - r + q))
 \end{aligned}$$

which gives (3.3b).

Since

$$\mathbf{E}(z_{12}^{aa} z_{12}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{21}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{13}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{31}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{23}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{32}^{bb}) = 0,$$

result in (C.19) yields

$$\rho(aa, bb) = -2p_a^2 p_b^2 (k^2 - r + q). \quad (\text{C.20})$$

Since

$$\mathbf{E}(z_{12}^{aa} z_{12}^{ab}) = \mathbf{E}(z_{12}^{aa} z_{21}^{ab}) = \mathbf{E}(z_{12}^{aa} z_{31}^{ab}) = \mathbf{E}(z_{12}^{aa} z_{32}^{ab}) = 0$$

and

$$\mathbf{E}(z_{12}^{aa} z_{13}^{ab}) = \mathbf{E}(z_{12}^{aa} z_{23}^{ab}) = p_a^2 p_b,$$

by (C.19) we obtain

$$\begin{aligned} \rho(aa, ab) &= p_a^2 p_b k(k-1) + p_a^2 p_b (k^2 - 2r) - 2p_a^3 p_b (k^2 - r + q) \\ &= p_a^2 p_b (2k^2 - k - 2r - 2p_a(k^2 - r + q)). \end{aligned} \quad (\text{C.21})$$

As

$$\mathbf{E}(z_{12}^{aa} z_{12}^{ba}) = \mathbf{E}(z_{12}^{aa} z_{21}^{ba}) = \mathbf{E}(z_{12}^{aa} z_{13}^{ba}) = \mathbf{E}(z_{12}^{aa} z_{23}^{ba}) = 0$$

and

$$\mathbf{E}(z_{12}^{aa} z_{31}^{ba}) = \mathbf{E}(z_{12}^{aa} z_{32}^{ba}) = p_a^2 p_b,$$

by equation (C.19) we get

$$\begin{aligned} \rho(aa, ba) &= p_a^2 p_b (k^2 - 2r) + 2p_a^2 p_b q - 2p_a^3 p_b (k^2 - r + q) \\ &= p_a^2 p_b (k^2 - 2r + 2q - 2p_a(k^2 - r + q)). \end{aligned} \quad (\text{C.22})$$

As

$$\mathbf{E}(z_{12}^{ab} z_{21}^{ba}) = p_a p_b, \mathbf{E}(z_{12}^{ab} z_{31}^{ba}) = p_a p_b^2, \mathbf{E}(z_{12}^{ab} z_{23}^{ba}) = p_a^2 p_b$$

and

$$\mathbf{E}(z_{12}^{ab} z_{12}^{ba}) = \mathbf{E}(z_{12}^{ab} z_{13}^{ba}) = \mathbf{E}(z_{12}^{ab} z_{32}^{ba}) = 0,$$

equation in (C.19) gives

$$\begin{aligned}\rho(ab, ba) &= 2p_a p_b r + (p_a p_b^2 + p_a^2 p_b)(k^2 - 2r) - 2p_a^2 p_b^2(k^2 - r + q) \\ &= p_a p_b (2r + (p_a + p_b)(k^2 - 2r) - 2p_a p_b(k^2 - r + q))\end{aligned}\quad (C.23)$$

Since

$$\mathbf{E}(z_{12}^{aa} z_{12}^{bc}) = \mathbf{E}(z_{12}^{aa} z_{21}^{bc}) = \mathbf{E}(z_{12}^{aa} z_{13}^{bc}) = \mathbf{E}(z_{12}^{aa} z_{31}^{bc}) = \mathbf{E}(z_{12}^{aa} z_{23}^{bc}) = \mathbf{E}(z_{12}^{aa} z_{32}^{bc}) = 0,$$

by the result in (C.19) we get

$$\rho(aa, bc) = -2p_a^2 p_b p_c(k^2 - r + q). \quad (C.24)$$

As

$$\mathbf{E}(z_{12}^{ab} z_{12}^{ac}) = \mathbf{E}(z_{12}^{ab} z_{21}^{ac}) = \mathbf{E}(z_{12}^{ab} z_{31}^{ac}) = \mathbf{E}(z_{12}^{ab} z_{23}^{ac}) = \mathbf{E}(z_{12}^{ab} z_{32}^{ac}) = 0$$

and

$$\mathbf{E}(z_{12}^{ab} z_{13}^{ac}) = p_a p_b p_c,$$

by (C.19) we obtain

$$\begin{aligned}\rho(ab, ac) &= p_a p_b p_c k(k - 1) - 2p_a^2 p_b p_c(k^2 - r + q) \\ &= p_a p_b p_c(k(k - 1) - 2p_a(k^2 - r + q)).\end{aligned}\quad (C.25)$$

As

$$\mathbf{E}(z_{12}^{ab} z_{12}^{bc}) = \mathbf{E}(z_{12}^{ab} z_{21}^{bc}) = \mathbf{E}(z_{12}^{ab} z_{13}^{bc}) = \mathbf{E}(z_{12}^{ab} z_{31}^{bc}) = \mathbf{E}(z_{12}^{ab} z_{23}^{bc}) = 0$$

and

$$\mathbf{E}(z_{12}^{ab} z_{23}^{bc}) = p_a p_b p_c,$$

equation (C.19) yields

$$\begin{aligned}\rho(ab, bc) &= p_a p_b p_c(k^2 - 2r) - 2p_a p_b^2 p_c(k^2 - r + q) \\ &= p_a p_b p_c(k^2 - 2r - 2p_b(k^2 - r + q)).\end{aligned}\quad (C.26)$$

Since

$$\mathbf{E}(z_{12}^{ab} z_{12}^{cb}) = \mathbf{E}(z_{12}^{ab} z_{21}^{cb}) = \mathbf{E}(z_{12}^{ab} z_{13}^{cb}) = \mathbf{E}(z_{12}^{ab} z_{31}^{cb}) = \mathbf{E}(z_{12}^{ab} z_{23}^{cb}) = 0$$

and

$$\mathbf{E}(z_{12}^{ab} z_{32}^{cb}) = p_a p_b p_c,$$

result in (C.19) gives

$$\begin{aligned} \rho(ab, cb) &= 2p_a p_b p_c q - 2p_a p_b^2 p_c (k^2 - r + q) \\ &= 2p_a p_b p_c (q - p_b (k^2 - r + q)). \end{aligned} \quad (C.27)$$

Finally, since

$$\mathbf{E}(z_{12}^{ab} z_{12}^{cd}) = \mathbf{E}(z_{12}^{ab} z_{21}^{cd}) = \mathbf{E}(z_{12}^{ab} z_{13}^{cd}) = \mathbf{E}(z_{12}^{ab} z_{31}^{cd}) = \mathbf{E}(z_{12}^{ab} z_{23}^{cd}) = \mathbf{E}(z_{12}^{ab} z_{32}^{cd}) = 0,$$

by equation (C.19) we have

$$\rho(ab, cd) = -2p_a p_b p_c p_d (k^2 - r + q) \quad (C.28)$$

and the results in (C.20)-(C.28) imply the ones in (3.3c)-(3.3k), respectively.

## C.2 Under Type 2 RL

By following the same idea introduced in Appendix C.1, replacing  $s$  with  $n$  in the results in (C.5)-(C.14) gives

$$\begin{aligned} \mathbf{E}(N_{\alpha\beta} N_{\theta\gamma}) &= \mathbf{E}(z_{12}^{\alpha\beta} z_{12}^{\theta\gamma}) nk + \mathbf{E}(z_{12}^{\alpha\beta} z_{21}^{\theta\gamma}) 2R_n^{[k]} + \mathbf{E}(z_{12}^{\alpha\beta} z_{13}^{\theta\gamma}) nk(k-1) \\ &\quad + \mathbf{E}(z_{12}^{\alpha\beta} z_{31}^{\theta\gamma}) (nk^2 - 2R_n^{[k]}) + \mathbf{E}(z_{12}^{\alpha\beta} z_{23}^{\theta\gamma}) (nk^2 - 2R_n^{[k]}) \\ &\quad + \mathbf{E}(z_{12}^{\alpha\beta} z_{32}^{\theta\gamma}) 2Q_n^{[k]} + \mathbf{E}(z_{12}^{\alpha\beta} z_{34}^{\theta\gamma}) (n^2 k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \end{aligned} \quad (C.29)$$

for every  $n \geq \max\{4, k+1\}$ .

We first compute the variances. Consider  $\mathbf{Var}(N_{aa}) = \mathbf{E}(N_{aa}^2) - (\mathbf{E}(N_{aa}))^2$ . Since

$$\begin{aligned} \mathbf{E}(z_{12}^{aa} z_{12}^{aa}) &= \mathbf{E}(z_{12}^{aa} z_{21}^{aa}) = p_{aa}, \mathbf{E}(z_{12}^{aa} z_{34}^{aa}) = p_{aaaa} \\ \mathbf{E}(z_{12}^{aa} z_{13}^{aa}) &= \mathbf{E}(z_{12}^{aa} z_{31}^{aa}) = \mathbf{E}(z_{12}^{aa} z_{23}^{aa}) = \mathbf{E}(z_{12}^{aa} z_{32}^{aa}) = p_{aaa}, \end{aligned}$$

by (C.29) we obtain

$$\begin{aligned} \mathbf{Var}(N_{aa}) &= p_{aa}(nk + 2R_n^{[k]}) + p_{aaa}(3nk^2 - nk - 4R_n^{[k]} + 2Q_n^{[k]}) \\ &\quad + p_{aaaa}(n^2 k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - (nk p_{aa})^2 \\ &= p_{aa}(nk + 2R_n^{[k]}) + p_{aaa}(3nk^2 - nk - 4R_n^{[k]} + 2Q_n^{[k]}) \\ &\quad + p_{aaaa}(-3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) + n^2 k^2 (p_{aaaa} - p_{aa}^2). \end{aligned} \quad (C.30)$$

Now for  $a \neq b$ , consider  $\mathbf{Var}(N_{ab}) = \mathbf{E}(N_{ab}^2) - (\mathbf{E}(N_{ab}))^2$ . Since

$$\begin{aligned}\mathbf{E}(z_{12}^{ab} z_{12}^{ab}) &= p_{ab}, \mathbf{E}(z_{12}^{ab} z_{21}^{ab}) = 0, \mathbf{E}(z_{12}^{ab} z_{13}^{ab}) = p_{abb}, \\ \mathbf{E}(z_{12}^{ab} z_{31}^{ab}) &= \mathbf{E}(z_{12}^{ab} z_{23}^{ab}) = 0, \mathbf{E}(z_{12}^{ab} z_{32}^{ab}) = p_{aab}, \mathbf{E}(z_{12}^{ab} z_{34}^{ab}) = p_{aabb},\end{aligned}$$

equation (C.29) gives

$$\begin{aligned}\mathbf{Var}(N_{ab}) &= p_{ab}nk + p_{abb}nk(k-1) + 2p_{aab}Q_n^{[k]} \\ &\quad + p_{aabb}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - (nk p_{ab})^2 \\ &= p_{ab}nk + p_{abb}nk(k-1) + 2p_{aab}Q_n^{[k]} \\ &\quad + p_{aabb}(-3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) + n^2k^2(p_{aabb} - p_{ab}^2).\end{aligned}\quad (C.31)$$

As

$$\mathbf{E}(z_{12}^{aa} z_{12}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{21}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{13}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{31}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{23}^{bb}) = \mathbf{E}(z_{12}^{aa} z_{32}^{bb}) = 0$$

and

$$\mathbf{E}(z_{12}^{aa} z_{34}^{bb}) = p_{aabb},$$

equation (C.29) yields

$$\mathbf{Cov}(N_{aa}, N_{bb}) = p_{aabb}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{aa}p_{bb}n^2k^2. \quad (C.32)$$

Since

$$\begin{aligned}\mathbf{E}(z_{12}^{aa} z_{12}^{ab}) &= \mathbf{E}(z_{12}^{aa} z_{21}^{ab}) = 0, \mathbf{E}(z_{12}^{aa} z_{13}^{ab}) = p_{aab}, \mathbf{E}(z_{12}^{aa} z_{31}^{ab}) = 0, \\ \mathbf{E}(z_{12}^{aa} z_{23}^{ab}) &= p_{aab}, \mathbf{E}(z_{12}^{aa} z_{32}^{ab}) = 0, \mathbf{E}(z_{12}^{aa} z_{34}^{ab}) = p_{aaab},\end{aligned}$$

by the result in (C.29) we obtain

$$\begin{aligned}\mathbf{Cov}(N_{aa}, N_{ab}) &= p_{aab}(2nk^2 - nk - 2R_n^{[k]}) + p_{aaab}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{aa}p_{ab}n^2k^2.\end{aligned}\quad (C.33)$$

Since

$$\begin{aligned}\mathbf{E}(z_{12}^{aa} z_{12}^{ba}) &= \mathbf{E}(z_{12}^{aa} z_{21}^{ba}) = \mathbf{E}(z_{12}^{aa} z_{13}^{ba}) = 0, \mathbf{E}(z_{12}^{aa} z_{31}^{ba}) = p_{aab}, \\ \mathbf{E}(z_{12}^{aa} z_{23}^{ba}) &= 0, \mathbf{E}(z_{12}^{aa} z_{32}^{ba}) = p_{aab}, \mathbf{E}(z_{12}^{aa} z_{34}^{ba}) = p_{aaab},\end{aligned}$$

by (C.29) we have

$$\begin{aligned} \mathbf{Cov}(N_{aa}, N_{ba}) &= p_{aab}(nk^2 - 2R_n^{[k]} + 2Q_n^{[k]}) + p_{aaab}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{aa}p_{ab}n^2k^2. \end{aligned} \quad (C.34)$$

As

$$\begin{aligned} \mathbf{E}(z_{12}^{ab}z_{12}^{ba}) &= 0, \mathbf{E}(z_{12}^{ab}z_{21}^{ba}) = p_{ab}, \mathbf{E}(z_{12}^{ab}z_{13}^{ba}) = 0, \mathbf{E}(z_{12}^{ab}z_{31}^{ba}) = p_{abb}, \\ \mathbf{E}(z_{12}^{ab}z_{23}^{ba}) &= p_{aab}, \mathbf{E}(z_{12}^{ab}z_{32}^{ba}) = 0, \mathbf{E}(z_{12}^{ab}z_{34}^{ba}) = p_{aabb}, \end{aligned}$$

equation (C.29) gives

$$\begin{aligned} \mathbf{Cov}(N_{ab}, N_{ba}) &= 2p_{ab}R_n^{[k]} + (p_{abb} + p_{aab})(nk^2 - 2R_n^{[k]}) \\ &\quad + p_{aabb}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{ab}p_{ba}n^2k^2. \end{aligned} \quad (C.35)$$

Since

$$\mathbf{E}(z_{12}^{aa}z_{12}^{bc}) = \mathbf{E}(z_{12}^{aa}z_{21}^{bc}) = \mathbf{E}(z_{12}^{aa}z_{13}^{bc}) = \mathbf{E}(z_{12}^{aa}z_{31}^{bc}) = \mathbf{E}(z_{12}^{aa}z_{23}^{bc}) = \mathbf{E}(z_{12}^{aa}z_{32}^{bc}) = 0$$

and

$$\mathbf{E}(z_{12}^{aa}z_{34}^{bc}) = p_{aabc},$$

by the result in (C.29) we obtain

$$\mathbf{Cov}(N_{aa}, N_{bc}) = p_{aabc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{aa}p_{bc}n^2k^2. \quad (C.36)$$

Since

$$\begin{aligned} \mathbf{E}(z_{12}^{ab}z_{12}^{ac}) &= \mathbf{E}(z_{12}^{ab}z_{21}^{ac}) = 0, \mathbf{E}(z_{12}^{ab}z_{13}^{ac}) = p_{abc}, \\ \mathbf{E}(z_{12}^{ab}z_{31}^{ac}) &= \mathbf{E}(z_{12}^{ab}z_{23}^{ac}) = \mathbf{E}(z_{12}^{ab}z_{32}^{ac}) = 0, \mathbf{E}(z_{12}^{ab}z_{34}^{ac}) = p_{aabc}, \end{aligned}$$

equation (C.29) yields

$$\begin{aligned} \mathbf{Cov}(N_{ab}, N_{ac}) &= p_{abc}nk(k-1) + p_{aabc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{ab}p_{ac}n^2k^2. \end{aligned} \quad (C.37)$$

As

$$\begin{aligned} \mathbf{E}(z_{12}^{ab}z_{12}^{bc}) &= \mathbf{E}(z_{12}^{ab}z_{21}^{bc}) = \mathbf{E}(z_{12}^{ab}z_{13}^{bc}) = \mathbf{E}(z_{12}^{ab}z_{31}^{bc}) = 0, \\ \mathbf{E}(z_{12}^{ab}z_{23}^{bc}) &= p_{abc}, \mathbf{E}(z_{12}^{ab}z_{32}^{bc}) = 0, \mathbf{E}(z_{12}^{ab}z_{34}^{bc}) = p_{abbc}, \end{aligned}$$

by the result in (C.29) we have

$$\begin{aligned} \mathbf{Cov}(N_{ab}, N_{bc}) &= p_{abc}(nk^2 - 2R_n^{[k]}) + p_{abbc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) \\ &\quad - p_{ab}p_{bc}n^2k^2. \end{aligned} \quad (C.38)$$

Since

$$\begin{aligned} \mathbf{E}(z_{12}^{ab}z_{12}^{cb}) &= \mathbf{E}(z_{12}^{ab}z_{21}^{cb}) = \mathbf{E}(z_{12}^{ab}z_{13}^{cb}) = \mathbf{E}(z_{12}^{ab}z_{31}^{cb}) = \mathbf{E}(z_{12}^{ab}z_{23}^{cb}) = 0, \\ \mathbf{E}(z_{12}^{ab}z_{32}^{cb}) &= p_{abc}, \mathbf{E}(z_{12}^{ab}z_{34}^{cb}) = p_{abbc}, \end{aligned}$$

by the equation in (C.29) we get

$$\mathbf{Cov}(N_{ab}, N_{cb}) = 2p_{abc}Q_n^{[k]} + p_{abbc}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{ab}p_{cb}n^2k^2. \quad (C.39)$$

Finally, since

$$\mathbf{E}(z_{12}^{ab}z_{12}^{cd}) = \mathbf{E}(z_{12}^{ab}z_{21}^{cd}) = \mathbf{E}(z_{12}^{ab}z_{13}^{cd}) = \mathbf{E}(z_{12}^{ab}z_{31}^{cd}) = \mathbf{E}(z_{12}^{ab}z_{23}^{cd}) = \mathbf{E}(z_{12}^{ab}z_{32}^{cd}) = 0$$

and

$$\mathbf{E}(z_{12}^{ab}z_{34}^{cd}) = p_{abcd},$$

by the result in (C.29) we have

$$\mathbf{Cov}(N_{ab}, N_{cd}) = p_{abcd}(n^2k^2 - 3nk^2 + 2R_n^{[k]} - 2Q_n^{[k]}) - p_{ab}p_{cd}n^2k^2. \quad (C.40)$$

### C.3 Derivation of $\rho(\alpha\beta, \theta\gamma)$ under Type 2 RL

We next compute the values of  $\rho(\alpha\beta, \theta\gamma)$  under type 2 RL. Recall that we assume  $n_i/n \rightarrow p_i$  for every  $i \in \{1, \dots, m\}$ ,  $R_n^{[k]}/n \rightarrow r$  and  $Q_n^{[k]}/n \rightarrow q$  as  $n \rightarrow \infty$ . Also note that  $p_{\alpha\beta} \rightarrow p_{\alpha}p_{\beta}$ ,  $p_{\alpha\beta\theta} \rightarrow p_{\alpha}p_{\beta}p_{\theta}$  and  $p_{\alpha\beta\theta\gamma} \rightarrow p_{\alpha}p_{\beta}p_{\theta}p_{\gamma}$  as  $n$  goes to infinity. Note that

$$\begin{aligned} n(p_{aaaa} - p_{aa}^2) &= \frac{n_a(n_a - 1)(n_a - 2)(n_a - 3)}{(n - 1)(n - 2)(n - 3)} - n \left( \frac{n_a(n_a - 1)}{n(n - 1)} \right)^2 \\ &= \frac{n_a(n_a - 1)}{n - 1} \left( \frac{(n_a - 2)(n_a - 3)}{(n - 2)(n - 3)} - \frac{n_a(n_a - 1)}{n(n - 1)} \right) \\ &= \frac{n_a(n_a - 1)(-4n_an(n - n_a) + 6(n^2 - n_a^2 - n + n_a))}{n(n - 1)^2(n - 2)(n - 3)} \\ &\rightarrow -4p_a^3(1 - p_a), \end{aligned}$$

as  $n \rightarrow \infty$ , and hence, (C.30) implies

$$\begin{aligned}\sigma_{aa}^2 &= p_a^2(k+2r) + p_a^3(3k^2 - k - 4r + 2q) + p_a^4(-3k^2 + 2r - 2q) - 4p_a^3(1-p_a)k^2 \\ &= p_a^2(k+2r - p_a(k^2 + k + 4r - 2q) + p_a^2(k^2 + 2r - 2q)),\end{aligned}$$

so we have (3.17a). Notice that

$$\begin{aligned}n(p_{aabb} - p_{ab}^2) &= \frac{n_a(n_a-1)n_b(n_b-1)}{(n-1)(n-2)(n-3)} - n \left( \frac{n_a n_b}{n(n-1)} \right)^2 \\ &= \frac{n_a n_b}{n-1} \left( \frac{(n_a-1)(n_b-1)}{(n-2)(n-3)} - \frac{n_a n_b}{n(n-1)} \right) \\ &= \frac{n_a n_b (n(4n_a n_b - n n_a - n n_b) + n^2 + n n_a + n n_b - 6n_a n_b - n)}{n(n-1)^2(n-2)(n-3)} \\ &\rightarrow p_a p_b (4p_a p_b - p_a - p_b),\end{aligned}$$

as  $n \rightarrow \infty$ , and thus, (C.31) yields

$$\begin{aligned}\sigma_{ab}^2 &= p_a p_b k + p_a p_b^2 k(k-1) + 2p_a^2 p_b q + p_a^2 p_b^2 (-3k^2 + 2r - 2q) + p_a p_b (4p_a p_b - p_a - p_b) k^2 \\ &= p_a p_b (k - p_b k + p_a(2q - k^2) + p_a p_b (k^2 + 2r - 2q)),\end{aligned}$$

so we get (3.17b).

Similarly, it is easy to verify that

$$n(p_{aabb} - p_{aa} p_{bb}) \rightarrow 4p_a^2 p_b^2, \quad (C.41a)$$

$$n(p_{aaab} - p_{aa} p_{ab}) \rightarrow 2p_a^2 p_b (2p_a - 1), \quad (C.41b)$$

$$n(p_{aabc} - p_{aa} p_{bc}) \rightarrow 4p_a^2 p_b p_c, \quad (C.41c)$$

$$n(p_{aabc} - p_{ab} p_{ac}) \rightarrow p_a p_b p_c (4p_a - 1), \quad (C.41d)$$

$$n(p_{abbc} - p_{ab} p_{bc}) \rightarrow p_a p_b p_c (4p_b - 1), \quad (C.41e)$$

$$n(p_{abbc} - p_{ab} p_{cb}) \rightarrow p_a p_b p_c (4p_b - 1), \quad (C.41f)$$

$$n(p_{abcd} - p_{ab} p_{cd}) \rightarrow p_a p_b p_c p_d. \quad (C.41g)$$

Therefore, using the results in (C.41a)-(C.41g) and (C.32)-(C.40), one can easily derive the equations in (3.16c)-(3.16k).

## BIBLIOGRAPHY

- Albert, A. (1972). *Regression and the Moore-Penrose Pseudoinverse*. Academic Press, Inc., New York, NY.
- Avram, F. and Bertsimas, D. (1993). On central limit theorems in geometrical probability. *The Annals of Applied Probability*, 3(4):1033–1046.
- Bahadır, S. and Ceyhan, E. (2016). On the number of reflexive and shared nearest neighbor pairs in one-dimensional uniform data. Technical Report e-print arXiv:1605.01940.
- Baldi, P. and Rinott, Y. (1989). On normal approximations of distributions in terms of dependency graph. *The Annals of Probability*, 17(4):1646–1650.
- Barbour, A. D. and Chen, L. H. Y. (2005). *Stein’s Method and Applications, Lecture Notes Series, vol. No.5.*, pages 223–245. Institute for Mathematical Sciences, National University of Singapore, Singapore University Press and World Scientific, Singapore.
- Barbour, A. D. and Eagleson, G. K. (1986). Random association of symmetric arrays. *Stochastic Analysis and Applications*, 4(3):239–281.
- Beer, E., Fill, J. A., Janson, S., and Scheinerman, E. R. (2011). On vertex, edge, and vertex-edge random graphs. *The Electronic Journal of Combinatorics*, 18(1):P110, 20pp.
- Bickel, P. J. and Breiman, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *The Annals of Probability*, 11(1):185–214.
- Bloznelis, M. (2010). A random intersection digraph: Indegree and outdegree distributions. *Discrete Mathematics*, 310(19):2560–2566.
- Bloznelis, M., Gotze, F., and Jaworski, J. (2012). Birth of a strongly connected giant in an inhomogeneous random digraph. *Journal of Applied Probability*, 49(3):601–611.
- Brito, M. R., Quiroz, A. J., and Yukich, J. E. (2013). Intrinsic dimension identification via graph-theoretic methods. *Journal of Multivariate Analysis*, 116:263–277.

- Ceyhan, E. (2009). Overall and pairwise segregation tests based on nearest neighbor contingency tables. *Computational Statistics & Data Analysis*, 53(8):2786–2808.
- Ceyhan, E. (2011). Spatial clustering tests based on the domination number of a new random digraph family. *Communications in Statistics -Theory and Methods*, 40(8):1363–1395.
- Ceyhan, E. (2014). Testing spatial symmetry using contingency tables based on nearest neighbor relations. *The Scientific World Journal*, Article ID 698296.
- Chartrand, G. and Lesniak, L. (1996). *Graphs & Digraphs*. Chapman & Hall/CRC, Boca Raton, FL.
- Clark, P. J. and Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453.
- Clark, P. J. and Evans, F. C. (1955). On some aspects of spatial pattern in biological populations. *Science*, 121:397–398.
- Conway, J. H. and Sloane, N. J. A. (1988). *Sphere Packings, Lattices, and Groups*. Springer Verlag, New York, NY.
- Cox, T. F. (1981). Reflexive nearest neighbors. *Biometrics*, 37(2):367–369.
- Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(1):73–104.
- Dacey, M. F. (1960). The spacing of river towns. *Annals of the Association of American Geographers*, 50(1):59–61.
- Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 33(2):129–135.
- Dixon, P. M. (1994). Testing spatial segregation using a nearest-neighbor contingency table. *Ecology*, 75(7):1940–1948.
- Dixon, P. M. (2002). Nearest-neighbor contingency table analysis of spatial segregation for several species. *Ecoscience*, 9(2):142–151.
- Enns, E. G., Ehlers, P. F., and Misi, T. (1999). A cluster problem as defined by nearest neighbours. *The Canadian Journal of Statistics*, 27(4):843–851.

- Eppstein, D., Paterson, M. S., and Yao, F. F. (1997). On nearest-neighbor graphs. *Discrete & Computational Geometry*, 17(3):263–282.
- Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297.
- Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Magyar Tudományos Akadémia Matematikai Kutatóintézet Közleménye*, 5:17–61.
- Friedman, J. H. and Rafsky, L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics*, 11(2):377–391.
- Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30:1141–1144.
- Graham, A. J. and Pike, D. A. (2008). A note on thresholds and connectivity in random directed graphs. *Atlantic Electronic Journal of Mathematics*, 3(1):1–5.
- Henze, N. (1987). On the fraction of random points with specified nearest-neighbour interactions and degree of attraction. *Advances in Applied Probability*, 19(4):873–895.
- Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783.
- Houdré, C. and Restrepo, R. (2010). A probabilistic approach to the asymptotics of the length of the longest alternating subsequence. *Electronic Journal of Combinatorics*, 17(1):Research Paper 168, 19 pp.
- Janson, S., Oleszkiewicz, K., and Ruciński, A. (2004). Upper tails for subgraph counts in random graphs. *Israel Journal of Mathematics*, 142:61–92.
- Kabatiansky, G. A. and Levenshtein, V. I. (1978). Bounds for packings on a sphere and in space. *Problems of Information Transmission*, 14:1–17.
- Karp, R. M. (1990). The transitive closure of a random digraph. *Random Structures Algorithms*, 1(1):73–93.
- Kozakova, I., Meester, R., and Nanda, S. (2006). The size of components in continuum nearest-neighbor graphs. *The Annals of Probability*, 34(2):528–538.

- Krivelevich, M., Lubetzky, E., and Sudakov, B. (2013). Longest cycles in sparse random digraphs. *Random Structures Algorithms*, 43(1):1–15.
- Luczak, T. (1990). The phase transition in the evolution of random digraphs. *Journal of Graph Theory*, 14(2):217–223.
- Luczak, T. and Seierstad, T. G. (2009). The critical behavior of random digraphs. *Random Structures Algorithms*, 35(3):271–293.
- Moon, J. W. (1968). *Topics on Tournaments*. Holt, Reinhart and Winston, New York.
- Musin, O. R. (2008). The kissing number in four dimensions. *Annals of Mathematics. Second Series*, 168(1):1–32.
- Najim, C. A. and Russo, R. P. (2003). On the number of subgraphs of a specified form embedded in a random graph. *Methodology and Computing in Applied Probability*, 5:23–33.
- Newman, C. M., Rinott, Y., and Tversky, A. (1983). Nearest neighbors and voronoi regions in certain point processes. *Advances in Applied Probability*, 15(4):726–751.
- Nowicki, K. and Wierman, J. C. (1988). Subgraph counts in random graphs using incomplete  $U$ -statistics method. *Discrete Mathematics*, 72:299–310.
- Penrose, M. (2003). *Random Geometric Graphs*. Oxford University Press, New York, NY.
- Penrose, M. D. and Yukich, J. E. (2001). Central limit theorems for some graphs in computational geometry. *The Annals of Applied Probability*, 11(4):1005–1041.
- Penrose, M. D. and Yukich, J. E. (2002). Limit theory for random sequential packing and deposition. *The Annals of Applied Probability*, 12(1):272–301.
- Pham, D. T., Möcks, J., and Sroka, L. (1989). Asymptotic normality of double-indexed linear permutation statistics. *Annals of the Institute of Statistical Mathematics*, 41(3):415–427.
- Pickard, D. K. (1982). Isolated nearest neighbors. *Journal of Applied Probability*, 19(2):444–449.
- Pielou, E. C. (1961). Segregation and symmetry in two-species populations as studied by nearest-neighbor relations. *Ecology*, 49:255–269.

- Reed, M. and Simon, B. (1980). *Methods of Modern Mathematical Physics. I. Functional Analysis*. Academic Press, Inc., 2nd edition, San Diego, CA.
- Ripley, B. D. (1981). *Spatial Statistics*. Wiley, New York, NY.
- Roberts, F. D. K. (1969). Nearest neighbours in a poisson ensemble. *Biometrika*, 56(2):401–406.
- Romik, D. (2011). Local extrema in random permutations and the structure of longest alternating subsequences. In *Proceedings of the 23rd International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2011), Reykjavik, Iceland, Association of Discrete Mathematics and Theoretical Computer Science*, pp. 825–834.
- Ruciński, A. (1988). When are small subgraphs of a random graph normally distributed? *Probability Theory and Related Fields*, 78:1–10.
- Schilling, M. F. (1986a). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.
- Schilling, M. F. (1986b). Mutual and shared neighbor probabilities: Finite- and infinite-dimensional results. *Advances in Applied Probability*, 18(2):388–405.
- Shang, Y. (2010). Laws of large numbers of subgraphs in directed random geometric networks. *International Electronic Journal of Pure and Applied Mathematics*, 2(2):69–79.
- Stanley, R. P. (2008). Longest alternating subsequences of permutations. *Michigan Mathematical Journal*, 57:675–687.
- Subramanian, C. R. (2003). Finding induced acyclic subgraphs in random digraphs. *The Electronic Journal of Combinatorics*, 10:#R46, 6pp.
- Wade, A. R. (2007). Explicit laws of large numbers for random nearest-neighbour-type graphs. *Advances in Applied Probability*, 39(2):326–342.
- Wyner, A. D. (1965). Capabilities of bounded discrepancy decoding. *Bell Systems Technical Journal*, 44:1061–1122.
- Yao, Y. C. and Simons, G. (1996). A large-dimensional independent and identically distributed property for nearest neighbor counts in poisson processes. *The Annals of Applied Probability*, 6(2):561–571.

- Yu, C. W. (2009). Computing subgraph probability of random geometric graphs with applications in quantitative analysis of ad hoc networks. *IEEE Journal on Selected Areas in Communications*, 27(7):1056–1065.
- Zhao, L., Bai, Z., Chao, C.-C., and Liang, W.-Q. (1997). Error bound in a central limit theorem of double-indexed permutation statistics. *The Annals of Statistics*, 25(5):2210–2227.

